DOCUMENT RESUME

ED 211 571 TM 810 907

AUTHOR	Kimmel, Wayne A.
TITLE	Putting Program Evaluation in Perspective for State and Local Government. Human Services Monograph Series, Number 18.
INSTITUTION	Department of Health and Human Services, Rockville, Md. Project Share.
SPONS AGENCY	Aspen Systems Corp., Germantown, Md.
PUB DATE	Apr 81
NOTE	66p.
AVAILABLE FROM	Project SHARE, P.O. Box 2309, Rockville, MD 20852 (free).
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Decision Making; Evaluation Methods; *Human Services; *Local Government; *Program Evaluation; *Public Agencies; *State Government
IDENTIFIERS	*Evaluation Utilization; Project SHARE

ABSTRACT

This paper attempts to put program evaluation into a broad but realistic perspective for managers and program practitioners in public agency human services programs at the state and local levels. It attempts to characterize the meaning and intent of program evaluation; to compare it with similar or allied approaches to improving decision making and management through the use of formal tools; to present some of the evidence on what difference evaluation seems to make in practice; and to discuss some of the basic issues raised by attempts to apply formal methods of evaluation. In conclusion, suggestions are offered to help the public agency practitioner decide whether or not to conduct formal program evaluation in a specific instance and what issues appear useful to consider. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Human Services

NUMBER 18 APRIL 1981 MONOGRAPH SERIES

ED211571

Putting Program Evaluation in Perspective for State and Local Government

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J Kath

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

PROJECT SHARE

A National Clearinghouse for Improving the Management of Human Services

Im 810 907

Human Services

NUMBER 18 APRIL 1981 MONOGRAPH SERIES

Putting Program Evaluation in Perspective for State and Local Government

Wayne A. Kimmel

PROJECT SHARE

A National Clearinghouse for Improving the Management of Human Services

Project SHARE has contracted for the preparation of a monograph series in order to survey the state of the knowledge or the state of the literature in selected subject areas of importance to the human services community. The monograph series provides an opportunity for authors to offer their views and opinions on these topics. It is the aim of Project SHARE to stimulate discussion through the publication of these monographs.

This monograph was prepared in fulfillment of a contract with Aspen Systems Corporation, publisher, as a contribution to the *Human Services Monograph Series*, Project SHARE, Department of Health and Human Services, Washington, DC. The views and opinions expressed in this monograph are entirely those of the author and are not necessarily those of DHHS or Aspen Systems Corporation.

Contents

Acknowledgments v

Preface vii

I. Introduction 1

Why Put Program Evaluation in Perspective? 1
A Brief History of Formal Program Evaluation 2
What Are the Claims for Formal Program Evaluation? 3

II. What Is Program Evaluation? 7

Is Evaluation Something New? 7
What Did Public Agencies Do Before Formal Program Evaluation? 8
The Emphasis of Program Evaluation 10
Relationships Among the Approaches 13
What Methods Are Proposed for Program Evaluation? 13

III. Selected Issues Raised by Formal Program Evaluation 17

The Role of Expectations 17
Values in Evaluation 18
The Criteria and Indicators Problem 20
Evaluation and "Causation" 22
Local Program Adaptations 23
Limits on Public Agency Control 24
Is Program Evaluation Research and/or Science? 25
Summary and Conclusions 25

IV. What Difference Does Program Evaluation Make in Practice? 29

Summaries of Selected Studies 30
General Conclusion 41
Intergovernmental Lessons 41

V. Proposed Reforms of Traditional Formal Evaluation 43

Sustain a Reasonable Measure of Self-Evaluation 43
Support Competitive Evaluations 43
Improve Citizen and Client Participation in Program Evaluation 44
Re-Examine Traditional Evaluation Premises 44
Conduct a Pre-Evaluation or Feasibility Assessment 44

VI. Evaluating the Expected Value of Doing a Formal Evaluation 47

Why Carry Out a Formal Evaluation Activity? 47
General Guidance for Program Evaluation 48
Pre-Evaluation Preparations (Homework) 50
Useful Practices and Rules of Thumb 52

References and Selected Bibliography 55

Acknowledgments

This monograph was prepared under contract with Project SHARE, an information clearinghouse operated by Aspen Systems Corporation, Rockville, Md., and funded by the Office of the Assistant Secretary for Planning and Evaluation (OASPE), Department of Health and Human Services (DHHS). Many individuals have assisted in the development of this paper. Eileen Wolff, DHHS project officer for SHARE, patiently provided suggestions and support. A number of people at Aspen Systems Corporation provided assistance. Anne Mehringer, SHARE project director, and Timothy Reed, SHARE project manager, placed SHARE's searching and archival resources at my disposal, and Ruth Curry, information specialist, cheerfully and speedily conducted several searches of the SHARE data base. Mehringer and Reed also arranged searches of the holdings of the data bases of both the Department of Housing and Urban Development's HUD USER and the Department of Justice's National Criminal Justice Reference Service clearinghouses. Carolyn Solomon, of the DHHS Evaluation Documentation Center in OASPE, cheerfully and quickly provided an abstract search and found many studies and reports which would be otherwise inaccessible.

Other individuals who provided interviews, documentation, ideas and assistance include Walton Francis, Director of the DHHS Office of Policy Analysis in OASPE, who encouraged the paper in the first place and contributed his sharp-minded and engaging discussion of the issues involved; Joseph Wholey, former DHHS Deputy Assistant Secretary for Evaluation; Richard Schmidt, DHHS Acting Deputy Assistant Secretary for Evaluation; Robert Raymond, Director for Intergovernmental Policy, DHHS; Alfred Schainblatt and Michael Gorman of the Urban Institute's State and Local Government Research Program; James W. Stockdill, Director of Planning, Evaluation and Legislation for the Health Resources Administration; Charles Windle, National Institute of Mental Health; Thomas L. Thorson, Professor of Political Science, Indiana University (South Bend); Mark Abramson, DHHS Office of Evaluation in OASPE; Edwin W. Zedlewski, Office of Research and Evaluation Methods, National Institute of Justice, Department of Justice; and dozens of State and local officials with whom I have talked over the years.

I thank several individuals who provided constructive criticism and suggestions on a draft: Francis; Schainblatt; Windle; Jane E. Fullarton, Consultant on Health, University of California, Berkeley; Henry Foley, Administrator, DHHS Health Resources Administration; John E. Glasson, Associate Dean for Budget and Administration, Western Reserve College, Case Western Reserve University; Thomas R. Vischi, Director of the Office of Treatment Policy of the DHHS Alcohol, Drug Abuse, and Mental Health Administration; and Evelyn Allin, Aspen Systems supervising editor for Project SHARE.

I am most grateful to my wife, Terri, who endured with understanding the sprawl of working papers and my recurrent preoccupations as this monograph inched to completion.

Twenty years ago I did some graduate study with Charles E. Lindblom at Yale University. His work, most recently *Usable Knowledge* (Lindblom and Cohen, 1979), has been a continuing spur to my thoughts.

The opinions and conclusions expressed are those of the author and do not reflect the official views or policies of DHHS, Aspen Systems Corporation, Project SHARE or any of the individuals mentioned herein.

Preface

This paper attempts to put program evaluation into a broad but realistic perspective for managers and program practitioners in public agency human services programs at the State and local levels. The bulk of the current literature on evaluation is occupied with the presentation and discussion of *technical methods* which might be employed in program evaluation. By contrast, this paper is neither a manual nor a how-to-do-it guide. Many of these already exist. Instead, it is an attempt to characterize the meaning and intent of program evaluation; to compare it with similar or allied approaches to improving decision making and management through the use of formal tools; to present some of the evidence on what difference evaluation seems to make in practice; and to discuss some of the basic issues raised by attempts to apply formal methods of evaluation. It concludes with suggestions to help the public agency practitioner decide whether or not to conduct formal program evaluation in a specific instance and what issues appear useful to consider.

The paper rests on three interrelated premises: First, public programs are "evaluated" from a number of different vantage points and through several different mechanisms all of the time, even though the volume of "formal technical evaluation" may be small or absent. Second, the general evaluation logic of assessing the worth and value of programs is appealing in principle, but the conditions under which formal evaluation will pay off in this way are more limited than is generally assumed. Third, formal program evaluation costs time, energy and money and should be treated like any other valuable commodity—with care and prudence. Like any other use of public resources, this one ought to meet a basic test of reasonable payoff: In any specific application, will program evaluation be "worth it?"

The bulk of the early literature on formal program evaluation (in the late sixties and early seventies) was optimistic about its usefulness in most circumstances. The evidence examined here along with testimony and reports from the field in recent years suggests, however, that the record is spotty, that the results of only a small proportion of actual formal evaluations appear to have impact and that reappraisals and more modest expectations about program evaluation are now appropriate. The paper identifies some of the proposals to reform evaluation practice and closes with some suggestions which might help a public agency official think about evaluation in a realistic way.

I have drawn selectively on the extensive evaluation literature and on discussions with knowledgeable individuals. I have also drawn generously on my own experience over the past 20 years as a public agency employee (DHEW), a consultant to public agencies and a field evaluator of several public agency programs. In particular, I have drawn on ideas and material from work on output measurement in elementary and secondary education done for the National Center for Education Statistics, U.S. Office of Education, at Georgetown University in the early seventies; writing on needs assessment methods for the Office of Program Systems, OASPE, DHEW in the late seventies; a review of research on evaluation and other management methods for the National Science Foundation carried out at the Urban Institute in the mid-seventies; and the insights and lessons shared with me by public and private agency officials at Federal, State and local levels I have had the good fortune to encounter in my work as a consultant. The background study for this paper was completed early spring 1980.

Part I traces briefly the early history of formal program evaluation as I understand it and identifies the many ambitious claims which were (and still are) made for its value and payoff to a public agency.

Part II develops context for understanding formal evaluation. It recalls the many mechanisms which society and public agencies have for evaluative judgments about programs; accents the

dominant features of program evaluation by comparing them with those of program planning, policy analysis and needs assessment; and lists some of the *alternative* methods, mechanisms and processes available for program evaluation.

Part III discusses the following basic issues which are raised when formal program evaluation is attempted in practice: the role of our expectations and values; the mixing of technical and value considerations in the selection of evaluation criteria and indicators or measures; a few of the realities which inhibit the aspiration to establish the "causes" of program effects; the tendency of some major traditional evaluation methods to miss or mask the inevitable and essential adaptations of general program designs to highly variable local circumstances; the impact on evaluation of the limits on public agency control; and whether program evaluation is "science." Since the discussion to this point covers a lot of terrain, a brief midpoint summary of major conclusions is provided in the concluding section.

Part IV summarizes some of the sketchy evidence contained in ten individual sources on the impact which formal program evaluation appears to make in practice.

Part V states briefly a few of the many recent proposed reforms of traditional evaluation theory and practice.

Part VI provides advice to the State or local agency official or program practitioner who may be considering carrying out some formal evaluation activities.

Some readers may wonder why experiences of the Federal Government are referenced so often. The reason is because lessons from available Federal experience (a dozen years) may contribute to a useful intergovernmental learning experience. Some of these lessons are recited at the end of part IV.

The casual, busy or knowledgeable reader may not want to read every word. The major parts may be read independently. The "bare bones" are contained in part I (Introduction), part III (Summary and Conclusions), part IV (Intergovernmental Lessons) and the summary advice given in part VI.

I. Introduction

Why Put Program Evaluation in Perspective?

Program evaluation carried out by or on behalf of governments in this country is now "big business." The nation spends well over a quarter billion dollars a year funding "evaluation." Thousands of field evaluation studies have been conducted over the past 15 years. Hundreds of books, papers and articles have been written and many published to create an evaluation literature which is now vast. It contains scores of how-to-do-it manuals, several evaluation research handbooks, dozens of case studies, proposed methods, growing reviews, critiques and criticisms, numerous exhortations to do evaluation, occasional horror stories from those who have done it, advice on how to avoid common pitfalls and a growing number of proposed reforms.

Substantial promises have been made in the past to agency executives, legislators and the public about the value of systematic, formalized evaluation of public programs. Evaluation will, it has been claimed, generate "objective" information about the operations and impacts of programs, tell us where they are strong and weak, indicate what is working and what is not and thereby save us heartache, frustration and millions, if not billions, by improving the design, performance, management, efficiency, effectiveness and impact of public programs.

Since the mid-sixties public outlays for evaluation have grown steadily. "Earmarks" and "set-asides" for evaluation have been written into scores of Federal program authorizations. Numerous and sometimes large staff offices focusing on evaluation have been established at the Federal level in particular and at the State and local levels as well. A large contract business has developed. An evaluation profession is said to be in the making. New professional evaluation societies and specialized journals and reviews appear at the rate of about one a year. Colleges and universities regularly offer courses and sometimes degrees in program and policy evaluation and research. A growth industry has been built around Federal mandates for evaluation. It represents a growing political constituency.

Yet despite this growth and prosperity (or partly because of it), major problems have emerged about the relevance, conduct and value of evaluation activities. Arguments are common concerning what evaluation really is or should be. Some claim we can all do it, others that only social scientists can do it and still others that we have not done it right yet. Some suggest that evaluation is common sense and others that it is science. Critics charge that the early promises about the payoff of evaluation have been broken, while proponents claim that expectations have been too high and that more time is needed to perfect evaluation tools. Some claim that evaluation has failed, that success is small and waste is frequent. Others claim that "knowledge is power" and that evaluation "research" is the only route to an "objective" basis for action. Others say that science has been perverted by politics and corrupted by contract economics. And recently, some leaders from the social sciences have taken critical looks at the excessive claims for and over-extended condition of social research applied to action programs.

In the meantime, many Federal and some State programs mandate, finance and encourage program evaluation as a part of program implementation and as a precondition for continued government support. The responsibility for a significant share of federally mandated and State-legislated evaluation falls to State and local government agencies. In the face of the ambiguity which surrounds the promise and payoff of formal evaluation, it is worthwhile to examine briefly the origins of the emphasis on formal evaluation, what evaluation seems to be and some of the major issues which it has raised.

A Brief History of Formal Program Evaluation

Once formal program evaluation had been mandated and practiced widely, scholars and commentators began to search for its roots. When did formal program evaluation begin? Lindblom and Cohen (1979) note the rise of what they call "professional social inquiry" (which includes evaluation) around the beginning of the century. Suchman (1967) reports that a concern with evaluation dates from the early beginnings of various fields of public service. He cites, for example, "a period of mounting health surveys and program evaluations" from 1907 to 1927, and the development of comparative community rating sheets by Chapin in 1914. He claims that it was not until after World War I that a "real demand for critical self-appraisal set in," but this demand appeared to result in "rather arbitrary, evaluation guides" and standards. (pp. 13-15) Freeman (1977) sees a sign for the current emphasis on formal program evaluation during the depression: "In 1935, an obscure sociologist, teaching at a then-small state university in the southern United States [Arkansas] published a paper pleading for the experimental evaluation of Franklin D. Roosevelt's New Deal Social Programs." (p. 18) These early antecedents, however, seem groping, sporadic and localized. Both Suchman and Freeman point most directly to the 1950's for the stirrings of a more general concern with and advocacy for systematic program evaluation. Though numerous, these efforts were still circumscribed, episodic and conducted in the mood of research and demonstration. It was not until the mid-sixties that the current wave of sustained modern program evaluation began. The Federal Government secured specific congressional authorization for program evaluation and began to invest sizable funds in it.

McLaughlin (1974), for example, reports that a requirement for the first significant evaluation of a major Federal program came in the provisions of the Elementary and Secondary Education Act (ESEA) of 1965. According to McLaughlin, Senator Robert Kennedy made his support for the new ESEA contingent on stronger reporting and executive oversight. In a meeting of principal drafters of the Act,

> Kennedy argued for an account of program activities as well as a strong USOE oversight role: "unless there is a meaningful program developed at the local level, which is really tested and checked by you [USOE], I don't think that this program is going to be effective."
>
> From this meeting there emerged the notions of a reporting and dissemination scheme that was subsequently included in the ESEA legislation, and of the evaluation provision that requires ESEA Title I projects to be regularly assessed for their effectiveness in meeting the "special educational needs of disadvantaged children." (p. 3)

It was also during this same period, in August 1965, that President Johnson ordered all the major Federal departments and agencies to install the so-called Planning-Programming-Budgeting System (PPBS) formerly used in the Department of Defense by Secretary Robert McNamara. That system was intended to improve the efficiency and effectiveness of resource allocation through systematic, multi-year program planning supported by systems, cost-benefit, cost-effectiveness and related analyses.

As analysts turned to the actual conduct of studies and analyses, however, they soon discovered to their surprise and disappointment that data on "output," "benefits" and "effectiveness" hardly existed. They also noticed that while many departments supported numerous and a wide array of program projects classified as "research and demonstrations," few of these yielded output, outcome, cost, benefit or effectiveness data which would support economic and "systems" modes of analysis. In DHEW, these discoveries precipitated a department-wide inventory by analysts in the newly created Office of Program Coordination (later Planning and Evaluation). Economists at the Bureau of the Budget (now OMB) and DHEW had noted that industry often spent about 4 to 6 percent on "research and development" activities. Why should a large department like DHEW not spend at least 1 to 2 percent on evaluation?

The inventory indicated that the funding of studies (which could be classified as studies of outcome or output) was uneven and low—below the level *judged* adequate. Based on the inventory, a desire to increase the volume of studies to support analysis, and with an arbitrary 1 percent figure in mind,

the Secretary issued a memorandum calling for a larger departmental investment in program analysis and evaluation. (It is noteworthy that the decision to initiate financial support for evaluation activities was not based on "verified" evidence that formal evaluation had an impact in practice. It is now clear that little or no formal evidence existed.) To generate funds for such analytical and evaluative work, the idea emerged to include in some new legislation an earmarked authorization for the Secretary to spend "directly or indirectly" (either through direct staff effort, grants or contracts) "up to 1%" of the annual appropriation for a program on "program evaluation."

The first authorization, of which I am aware, for the so-called "1% set aside" (which now supports the bulk of the DHHS evaluation effort) was included in the Department's proposed Partnership for Health Amendments of 1967. [Note: The Congressional Committee Report on the amendments contained several paragraphs drafted by the author, then a staff member of OASPE, DHEW, to justify the new authority.] Once the pattern was set, similar earmarked evaluation authorities were proposed for an ever-increasing number of social programs. In 1969, for example, a new administration decided to "blanket in" all the major departmental authorities for scores of human services programs under this evaluation authority.

In time, provisions in legislative statutes were written that not only authorized Federal-level evaluation but also authorized and frequently mandated program evaluation for recipients of Federal funds at the State and local levels as well. Thus began the Federal support for program evaluation which by 1976 had reached an estimated national level of over a quarter of a billion dollars. It is to the general problems and issues of formal program evaluation that this paper is devoted. We begin with a basic question.

What Are the Claims for Formal Program Evaluation?

According to many proponents, program evaluation:

1. Consists of the study of public programs and their impacts through the use of *systematic* (sometimes characterized as "scientific") *methods* of investigation and "research;"
2. Will generate a body of *valid, reliable* and presumably *verified* propositions and conclusions about the impacts and operations of programs;
3. Will thereby indicate in an *impartial* way what effects actually result from a given program and which features of a program account for them (their *causes*); and
4. Will make this *new* and "objective" information available to decision makers responsible for the program as a basis (partial but crucial) for deciding whether and to what extent a program is working, why it is working the way it is and, presumably, what can be done to improve it.

The application of evaluation methods and the provision of new unbiased information to decision makers will, it is argued, lead to an improved understanding of the program. This will lead, in turn, to an improvement in the overall "rationality" of decision making. The result of these improvements will be increased efficiency and effectiveness. Programs will work better. Resources will be saved. Public agencies will be held "accountable." The general welfare will be better served.

With these hopeful prospects in mind, early proponents of evaluation heralded two additional broad-based claims and assumptions:

1. Programs could be evaluated in their totality (later called "summative evaluation") to yield comprehensive information, conclusions and judgments on their overall worth and workability; and
2. Program evaluation could (and should) be applied to every significant program — the principle of Evaluation Universality.

The early claims for formalized evaluation were often made with vigor, assertiveness and an occasional touch of arrogance. Some viewed public agencies as largely entrenched, self-serving and lethargic. Decision makers and program managers were sometimes depicted as narrow-minded, myopic bureaucrats who did not appreciate the power of the tools of formal evaluation. They appeared to spend most of their time protecting their turf and covering their mistakes, while the "big issues" of refined program objectives, effectiveness measures and "causation" went by the wayside. Although decision makers presumably had power, they seemed to spend it in fights over office space and carpeting and not over dubious program premises and poor program designs.

By contrast, program evaluation was seen as a swift and sure-footed route to clearer objectives, reliable and impartial data, scientific "facts" and verified conclusions which could be used to root out ignorance, motivate bureaucrats, "depoliticize" decision making and, as Wildavsky (1979) has put it, speak truth to power. If a revolution were not in the offing, we seemed at least on the verge of a new era of rationality and reform. Evaluators aided by social science methods would set us free from self-interested politics. Or so it seemed.

These are, obviously, ambitious claims for formal program evaluation. To what extent have they been fulfilled? A growing body of evidence, critical review, reports from the field and self-criticism suggest that in terms of nearly all the early claims, formal program evaluation has fallen very short. Here is recent testimony from experienced evaluators.

- After 5 years experience at the Urban Institute with the theory and the application of formal program evaluation to a range of Federal programs, Schmidt, Scanlon and Bell (1979) opened their proposed reform of evaluation ("evaluability assessment") with this judgment:

  > ...Congress and the executive branch have increasingly invested in program evaluation over the past decade. Starting from nearly nothing in the early sixties, investment in evaluation grew to around a quarter of a billion dollars by 1976. Unfortunately, however, the investment has not yet paid off. Program evaluation has not led to successful policies or programs. Instead, it has been planned and implemented in isolation from Federal decisionmaking, and has produced little information of interest and utility to policymakers and managers. (p. 1)

- Rossi and his colleagues screened "several hundred" Federal evaluation RFP's (requests for proposal) "searching for examples we could use for didactic exercises in the Summer Institute. We were able to find less than a dozen that we could use..." A further search of "more than a hundred" completed evaluation research reports using "minimal" standards yielded not more than a half dozen of "high quality." Rossi concluded:

  > The fact of the matter is that most evaluations are still not worth much more than no evaluation at all. (Datta and Perloff, 1979, pp. 20-21)

- In their study of education evaluation, Alkin, Daillak and White (1979) "discern the very few discordant cries" that evaluation works and report:

  > In fact, the literature is replete with gloomy statements about the impotence or futility of evaluation. There seems to be a consensus in the literature that there has been little impact of evaluative research on program decisionmaking. (p. 14)

- Evaluators at the Rand Corporation recently reflected on dominant evaluation practices in education. In a set of engaging papers, they comment on the modest contribution of formal evaluation and propose several reforms to current practice. Editor John Pincus notes that:

  > Most policymakers want their programs to succeed; but most "scientific" evaluations address effects and indicate that student outcomes as measured by test scores, drop-out rates, and other such measures appear to be little affected by new government education programs. Such reports of "no significant effect" are generally unaccompanied by useful recommendations for program improvement or policy change. Meanwhile, policymakers seek to know not only about effects, but also about what is going on in the program.... In effect, what can result is a "dialog of the deaf," in which neither party understands the other's premises. Is it possible to reduce these tensions and improve the utility of evaluation to public policy? (Pincus, 1980, pp. 1-2)

The four Rand essays "find fault with current evaluation methods, each from a different perspective, and call for improvements" and "for a retreat from the somewhat over-ambitious pretensions of social science in the earlier years of evaluation studies. . . ." (p. 5)

- In her reflections on "Evaluation and Alchemy," McLaughlin (1980) of Rand recounts that the central Office of Education budget for evaluation "mushroomed" from about $1.2 million in 1968 to about $21 million in 1977:

  > But despite the energy and resources devoted to the task, many researchers and practitioners believe these evaluation efforts are largely a waste of time and money. (p. 41)

  She continues:

  > Our research supports the charge that much of the present evaluation is irrelevant and inappropriate—that most evaluations ask the wrong questions and use the wrong measures. (p. 42)

  Her proposed reforms run deep:

  > What we know about the process of change implies that evaluation models derived from other realities — microeconomics, medicine, and social psychology — simply do not fit the reality of a public social service system, education in particular. The logic of inquiry is wrong. And preoccupation with scientism and with fixing our traditional evaluation paradigms scants what we do know. . . . One major challenge for evaluators, then, is epistemological: to develop new and valid ways of knowing. (p. 46)

[Note: Epistemology refers to a branch of philosophy that investigates the nature and origin of knowledge. How do we get it and what does it rest on?]

- In a broad and self-conscious appraisal of "evaluation research," Campbell (1979), a widely quoted applied social scientist and methodologist, reports:

  > We cannot yet promise a set of professional skills guaranteed to make an important difference. In the few success stories of beneficial programs unequivocally evaluated, society has gotten by, or could have gotten by, without our help. We still lack instances of important contributions to societal innovation which were abetted by our methodological skills. The need for our specialty, and the specific recommendations we make, must still be justified by promise rather than by past performance. (p. 68)

The testimony of these evaluators is sobering if not disturbing. Why has program evaluation made such a seemingly poor showing? We attempt here a preliminary understanding by continuing with another basic question.

II. What Is Program Evaluation?

Though this is a simple and straightforward question, it appears to have no ready answer. In one of the early attempts to discuss evaluation, for example, Suchman (1967) called attention a dozen years ago to the fact that "evaluation despite its widespread popularity is poorly defined and improperly used" (p. 27). Block and Richardson recently remarked that "No concept is so misused in social science as evaluation...." (1979, p. 9) And the claim has been made occasionally that there are as many definitions of evaluation as there are writers on the subject.

As often the case, the dictionary offers a good first approximation to the meaning of evaluation. According to the New College Edition of the *American Heritage Dictionary of the English Language* (Morris, ed., 1976), "evaluate" means

> 1. To ascertain or fix the value or worth of. 2. To examine and judge; appraise; estimate: *"Plato has been evaluated as having one of the finest minds the world has produced"* (S.E. Frost, Jr.). 3. *Mathematics.* To calculate or set down the numerical value of; express numerically. — See synonyms at **estimate**....

At bottom, most writers on evaluation endorse the general dictionary notion: *to ascertain or fix the value or worth of* something.

Weiss (1972) captures the same idea:

> Evaluation is an elastic word that stretches to cover judgments of many kinds.... What all the uses of the word have in common is the notion of judging merit. Someone is examining and weighing a phenomenon... against some explicit or implicit yardstick. (p. 1)

The crux of the evaluation problem, then, appears to be establishing the worth or value or merit of something. All evaluations must start with this concern and inevitably return to it.

Although the word "program" is widely and commonly used in public agencies, it, too, has a variable meaning. Nearly any activity may be called a program: a research program, a regulatory program, a technical assistance program, a monitoring program, an audit program, a grant program and so on. In an evaluation context, it appears useful to use the term "program" in a general way to refer to *the organized use of public resources directed toward the accomplishment or achievement of one or more purposes and/or objectives.* The program evaluation of concern here is directed primarily to *human services programs:* child day care services, community mental health services, education services for the disadvantaged, manpower training and placement services, vocational rehabilitation and a wide variety of other health, welfare, education, housing and social services.

If we combine the basic meanings of the two words, "program evaluation" can be roughly characterized as:

> attempts to ascertain or fix the value or worth of the use of organized public resources directed toward one or more purposes and/or objectives.

Before we examine program evaluation further, a brief digression on the origins of evaluation will be useful.

Is Evaluation Something New?

In terms of its basic meaning of judging worth and value, human beings have probably always been engaged in evaluation. From time to time and sometimes continuously we evaluate, wittingly or not,

most aspects of our lives: our jobs, living circumstances, the behavior of our children, the things we buy and use, the weather, the news, art, political events, our health, taxes and the like. There runs through all we think, do and say a deep-seated, irresistible, and wholly human element of estimating and judging what has been and is happening, whether or not we like it and how much or how little. Both our everyday and professional languages are laden with evaluative and judgmental words, phrases, content and overtone. It is only a slight exaggeration to say that in our thoughts, choices and other behavior we are an "evaluating species" — well practiced at judging worth and value.

As individuals, however, we are, in principle, at liberty to assess the worth or value of something in light of our own private values and preferences. We do not usually have to take into account the values and preferences of others. As we move from strictly individual, private evaluation to making judgments on behalf of a family, a group, an agency, a community or a State, the scope of relevant values expands. In the arena of public policy making in a democratic political system, for example, we expect decision makers to take into account and reconcile the multiple and diverse values of those who have a stake in a policy or program, whether they are existing or potential clients, service providers, other agency officials, the public (defined in some way as taxpayers, beneficiaries, etc.) or other "stakeholders" and "interested parties."

The problem of assessing worth or merit takes on a complex character as we move from *individual* to *collective* or social values as the basis for judgment. As we note later, some of the dilemma and difficulties of program evaluation undertaken on behalf of public agencies arise from the attempt to move from evaluation at the *individual* level to evaluation at a *collective* level.

Before identifying the major methods by which *formal* social and program evaluation is to occur, it will be worthwhile to ask another basic question.

What Did Public Agencies Do Before Formal Program Evaluation?

If some of the criticisms and claims of early advocates of formal evaluation were taken at face value, one might conclude that before the advent of formal program evaluation public agencies had no way of knowing how well existing programs were faring, that agency officials and program managers operated in a vacuum of information and knowledge about program workability and impact, that feedback did not exist or was fatally flawed and that decision makers merely "flew by the seat of their pants" or trusted only to their "gut reactions."

There are kernels of truth to some of these criticisms but they under estimate the wide variety of governmental, social, economic and political information-generating, feedback and program-testing mechanisms which do exist. Here is a list of some of the familiar mechanisms, linked in practice by complex social, political and bureaucratic processes through which judgments about the value and worth of public agency programs and services are regularly rendered.

Official Public Mechanisms

Legislative/Council Review

Program authorization hearings and debates;
Budget review, hearings and debates;
Public hearings; and
Oversight hearings (or studies).

Executive Review

Program and budget review;
Special studies (e.g., study group, task force, commission, etc.);

Site visits and reports;
Audits (financial and sometimes program); and
Organizational and management analyses.

Judicial Review

Court rulings on issues of eligibility, rights, due process, etc.

General Public Review

Candidate selection through primaries and elections;

Communication between elected officials and constituents;

Consumer and client complaints and grievances; and

Media reporting (general and investigative).

Professional Review

Professional contribution to program design and implementation;

Testimony and reports by program and problem experts;

Criticism and commentary by social scientists; and

Policy statements and commentary by professional associations.

Special Interest Review

Case-stating, criticism, commentary, evidence-reporting and "pressure" by organized interest groups.

Review Through Market-Like Mechanisms

Competition among claimants for social resources.

Economic Market Mechanisms

Valuation through exchange, competition and pricing.

The degree of visibility, institutionalization and effectiveness of the functions performed by these mechanisms do, of course, vary. They may range from highly articulated and distinctive mechanisms to those which are rudimentary, informal and episodic. In some instances, no identifiable mechanism may exist at all.

Some proponents of formal program evaluation find the assorted legislative, administrative, judicial, social, economic and political mechanisms (and related processes) for evaluation inadequate: vested with special interests, preoccupied with political considerations and bereft of adequate formal evidence on the basis of which informed ("rational") judgments could be made about program design, funding and redirection. These mechanisms remain, however, among the most dominant and widely used vehicles by which collective social judgments are expressed about the use of resources in public programs. And they are the primary mechanisms through which the results of formalized program evaluation must be used, if they are to be used at all.

Some proponents of formal evaluation recognize the constraints that existing political processes and mechanisms place on the utilization of formal evaluation results. They sometimes suggest reform of the institutions and processes of politics as a way to increase the use of evaluation. There is no question that political and administrative institutions and processes can be reformed in attempts to increase their accountability. In general, however, it is a common error to miss or dismiss the value of existing *social processes of interaction* as available mechanisms for collective program evaluation. (Lindblom, 1965; Lindblom and Cohen, 1979)

The Emphasis of Program Evaluation

Every major method and approach to "improving" the performance of an organization focuses on, accents or emphasizes some aspects of an organization's behavior, structure, functions or processes (e.g., Kimmel, Dougan and Hall, 1974). In most cases, a "new" approach tends to make more explicit, more formalized and more central a set of functions which are already being performed though perhaps in a more implicit, informal and less sustained way. Management by Objectives (MBO), for example, focuses on internal short-range management goal and objective setting. It is intended to induce joint objective setting between superiors and subordinates and thereby increase communication between them. Program Evaluation Review Technique (PERT) is intended to improve an organization's capacity for defining and relating tasks, for work scheduling and for determining optimal or critical paths through complex interrelated activities by estimating and comparing their time and/or cost requirements. Organizational Development (OD) is intended to improve organizational performance by improving employee self-consciousness and interpersonal relations.

The literature both states and implies that program evaluation is intended to "rationalize decision making," to provide valid information on the performance of a program, and thereby to improve decisions about the design, level of funding, operation and management of a program. Similar claims are made for several other methods and approaches urged for use in public agencies; for example, program planning, policy analysis and needs assessment. A brief comparison of program evaluation with these three approaches will help highlight distinguishing features of program evaluation. Table 1 displays selected major features of these approaches which are reflected in the philosophy and general logic set out in the literature. These are, obviously, "average" representations. In all cases there are large and small variations from writer to writer and from application to application. The comparative table suggests these highlights:

General Contrasts

In a very summary way the four approaches have these general orientations:

Program planning focuses on the use of future resources to achieve a set of tentatively established goals and objectives over a multi-year period. This approach typically rests on a comprehensive view of an agency's programs and on estimates of future conditions, costs and expected results of programs, some of which are yet to be formulated.

Program evaluation, by contrast, focuses on a program already formulated and operating. The attempt is not to forecast or *predict* the future but to *retrodict* the past—to identify, gauge, and judge the value of the results which the program has already generated. It addresses the basic questions: What difference has the program already made? Is the program worthwhile? Does it work? How might it be improved?

Policy analysis focuses primarily on an existing or likely policy problem, its structure, seeming causes, possible policy responses and a comparison of alternative responses in terms of their estimated costs and effects. Though a policy analysis may include consideration of an existing program, attention is not limited to any existing program alternative.

Needs assessment represents an attempt to identify and assess the types and extent of perceived, reported or inferred "needs" in a defined population group. Existing programs are relevant to this approach in attempts to identify what are perceived to be gaps in existing services.

It is clear that these four approaches overlap (see Kimmel, 1977, and Morrill and Francis, January 1979).

Table 1

A Selected Comparison of Four Proposed "Rational Aids" To Decision

	POLICY ANALYSIS	NEEDS ASSESSMENT
MAJOR CONCEPTUAL SOURCES	Economics Choice theory Decision theory	Eclectic (unclear): Psychology Social work Survey research
MAJOR CONCEPTS	Systems view of a problem Structure of a problem Causes of a problem Alternative responses to the problem Costs and benefits of alternatives Criteria for choosing among alternatives Constraints	Needs: Individual Community Met Unmet Assessment: Estimating Valuing Judging Gaps in service: Estimated needs juxtaposed to existing services
DOMINANT FOCAL POINTS	Existing or expected policy problem Specification of alternative responses to the problem Comparison of alternatives	Perceived or inferred needs of population groups: Community Population at risk Target population Service population
INTENDED USES	Analytical input to decision making about a problem Planning Program design Resource allocation Rationalize decision making	Planning Priority setting Resource allocation Rationalize decision making
MAJOR DATA SOURCES	Multiple and varied depending on the nature of the problem Emphasis on the use of existing studies, information and data	Opinions of experts and groups Field surveys Social indicators Demographic indicators Epidemiological studies Incidence and prevalence studies Secondary data analysis
ORGANIZATIONAL LOCATION	Staff office serving decision makers	Rarely discussed Often performed outside the agency

Making

PROGRAM PLANNING	PROGRAM EVALUATION
Business management theory Economics Decision theory Forecasting Planning theory	Social science field research: especially from psychology and sociology Some economics and engineering Statistics
Resource constraints Budget costs Policy goals and objectives Program alternatives Tradeoffs Input-output relationships Future uncertainty Multi-year forecasts: Conditions Costs Results	Program goals and objectives Program outcome, impact and results Criteria of outcome and impact (or other change) Measures or indicators of the criteria Comparisons of changes in the measures or indicators Comparisons with and without the program
The future: Mix of old and new problems New goals and objectives Estimated resource availability Possible alternative courses of program development Estimated outputs, outcomes, impacts	An existing operating program Changes which result from the program, especially among clients or problem conditions Judgments about program worth based on observed, measured and inferred changes Program performance
Development of multi-year plans Context for current decisions about the use of future resources Recommendations on current decisions Rationalize decision making	Feedback on the results of existing program Improve program management Increase efficiency and effectiveness of programs Improve program design Rationalize decision making
Multiple and varied Time series data Analytical studies Evaluation studies	Multiple and varied depending on the indicators and measures selected Frequent emphasis on new data collection
Staff office serving decision makers	Staff office serving decision makers Often performed outside the agency

Major Conceptual Sources

The major conceptual or philosophical sources of much formal program evaluation appear to be the field research components of several social sciences, principally psychology and sociology. Some concepts and methods from economics and engineering (effectiveness, input-output relationships) are also employed. Many of the concepts for policy analysis come from microeconomics, decision theories and theories of choice. The sources of the concepts of needs assessment are unclear and eclectic, though they are probably derived from psychology and social work. Those of program planning are eclectic: business management theory, economics, forecasting and formal planning theory.

Major Concepts

The major concepts of program evaluation include an emphasis on *program goals and objectives, measures or indicators* which are to be derived from those goals and objectives, *changes* which occur due to the operation of the program, *outcomes or impacts* of the program reflected in an appraisal of those changes, and a set of *judgments* about the value or worth of the program. While a consideration of the *costs* of a given program or one of its elements may or may not be part of an evaluation, the notion of resource constraints is central to policy analysis. In principal, both approaches (evaluation and analysis) attempt to estimate "net benefits" of a program and both attempt to establish some notions about "cause and effect." These considerations are normally absent, for example, from needs assessment approaches.

Dominant Focal Points

The dominant focal points of program evaluation are judgments about the value or worth of a program, about probable causes and about results, based on measured changes which can be attributed to the program. Those of program planning are estimates of resource requirements and of the expected costs and results of a future mix of programs directed toward some tentatively established goals and objectives. Needs assessment focuses on unmet needs of a defined population group. Policy analysis focuses on existing or expected policy problems, and on an explicit comparison of alternative responses to those problems, whether or not a program already exists.

Intended Uses

The intended uses of all the approaches are expressed in claims that they will "improve the rationality of decision making": about future plans and priorities in the case of program planning; about major program, resource and management decisions in the case of program evaluation; about decisions on policy issues in the case of policy analysis; and about future plans to fill unmet needs in the case of needs assessment.

Major Data Sources

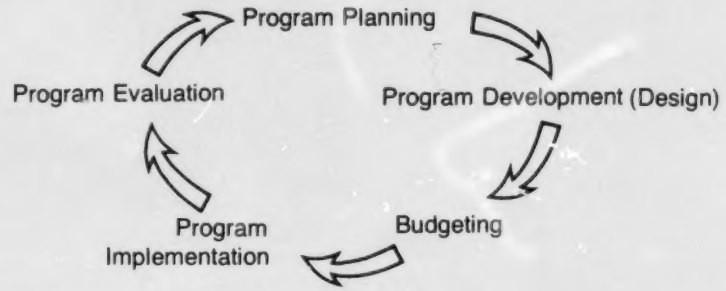
All the approaches appear to require a mixture of existing and new data. The approach of policy analysis usually includes an injunction to use existing data, studies and analysis creatively. The approach of needs assessment emphasizes new data collection, often through the use of surveys. Program evaluation focuses on the measurement of outcomes and impacts. Because existing data for these purposes are often scarce, the approach usually requires new field data collection, sometimes of an extensive variety.

Organizational Location

The prescriptions for program planning, program evaluation and policy analysis recommend that these activities be located in a staff office serving key decision makers. Needs assessments and program evaluation studies are, however, usually conducted by groups outside the agency.

Relationships Among the Approaches

Writers on program evaluation, like those on other formalized approaches which are urged on public agencies, often comment on the imperative or "logical" role of the formal approach in the affairs of an organization. Though represented in many alternative ways, a common diagram depicts several proposed management functions in a cyclical relationship something like this:



Here the approaches to decision making are presented in a closed loop of interdependent activities. In this cyclical sequence, one function leads directly to another. Plans are first developed, then programs are designed. Approved programs are then funded through budgeting and then implemented. Once operating, a program is evaluated and performance information is fed back to decision makers for use in yet another cycle of functions.

This representation is, obviously, an *ideal* model. It is derived not from empirical observations of or operations of real organizations but from an idealized/technical style of thinking based on a variety of "rationality" (there are many varieties) which has this step-wise form:

First, identify goals and objectives (Plan).
Second, specify alternatives to reach these objectives (Programs).
Third, compare the alternatives (Analysis).
Fourth, select the best one (Choice).
Fifth, fund the chosen alternative (Budget).
Sixth, set the program into operation (Implementation).
Seventh, assess the program in terms of its results (Evaluation).
Eighth, repeat the cycle.

While it has the appeal of simplicity, this rational, sequential model is rarely, if ever, followed in actual public agency practice. There are several reasons. Many of them spring from the basic nature and role of a public agency in a changing environment of social and political interaction. (Lindblom, 1965; Pressman and Wildavsky, 1973)

What Methods Are Proposed for Program Evaluation?

When earmarked funding for evaluation purposes first began in the Department of Health, Education, and Welfare, for example, it was assumed by many that the approaches and methods which would be employed in program evaluation studies would range in type and variety depending on the nature of the program to be evaluated. Case studies, surveys, field interviews, self-evaluation and informed observation and analysis had long-standing use. Critics argued that many of these techniques were weak tools for evaluation because they did not ensure adequate "objectivity," were sometimes used in an "unsystematic" way and did not always provide "reliable" data on the basis of which the "causes" (of, say, a change in student learning progress) could be established.

Other formal methods were also already in use in some government settings. Cost-benefit analysis, for example, had been applied in a number of areas including the comparison of alternative disease control programs. Methods of cost-effectiveness and systems analysis were widely discussed and applied with varying degrees of success. As we noted earlier, however, the application of these techniques frequently relied on the use of quantitative data which often did not exist. Critics of these methods also claimed that they placed excessive emphasis on the economic aspects of programs and not enough on changes in the program client (social and psychological), on the impacts which a program might have on an institution (a school or hospital) or on the wider community (mental health or the frequency of delinquency).

Social scientists, especially psychologists and sociologists who claimed a tradition of research, argued that the methods of field evaluation should come from the social sciences. There arose rather quickly what has since become a long and continuing discussion, debate and argument among interested parties over what *methods and approaches* were the more appropriate, reliable and preferred for program evaluation purposes. In his early and widely read book—Campbell (1979) calls it "the founding book"—*Evaluative Research*, for example, Suchman (1967) stated that he wanted to "retain the term 'evaluation' in its most common sense usage as referring to the general process of assessment or appraisal of value." (p. 7) When he turned to his main subject, however, he introduced this special condition:

> Thus from the beginning we would like to make it clear that we do not view the field of evaluation as having any methodology different from the scientific method; evaluation research is, first and foremost, *research* and as such must adhere as closely as possible to currently accepted standards of research methodology. . . ultimately the significance of the results must be determined according to the same scientific standards used to judge nonevaluative research. (p. 12)

Four chapters later, however, Suchman admitted:

> Examples of evaluative research which satisfy even the most elementary tenets of the scientific method are few and far between. (p. 74)

In 1970 in another early, influential and often-referenced book on evaluation, Wholey and his associates (1976) speak of "formal, organized evaluation" this way:

> In this sense evaluation is research, the application of the scientific method to experience with public programs to learn what happens as a result of program activities. (p. 19)

Caro (revised edition 1977), another widely read author, opened his edited collection of writings on evaluation, in the early seventies, with a general definition and then added his own emphasis on formal *research*:

> Program evaluation has two essential dimensions, one concerned with judgment and the other with information. Programs are conducted to achieve a goal, end, or outcome that is valued. Program evaluation produces judgments regarding the degree to which desired outcomes have been achieved or can be achieved. It leads to conclusions regarding the worth of organized efforts. Information is of critical importance in the evaluation process. Performance as known through verifiable procedures is related or contrasted to goals. The method through which such information is obtained is often a central point in evaluation. (p. 3)

Caro then mentions several alternative methods of evaluation: those we use in everyday life; accreditation, for example, through licensing; and cost analysis. He, like Suchman, then opts for evaluation *research*.

> Evaluation research may be considered a third tradition that is distinguished by its central concern for outcomes of treatment. It attempts to determine whether changes sought through an intervention actually come about. Further, evaluation research is concerned with the question of whether observed changes can reasonably be attributed to the intervention. Evaluation research, therefore, makes use not only of scientific method but procedures designed to test for causal connections. It is the evaluation research approach with which this volume is primarily concerned. (p. 5)

Suchman, Caro and other social scientists (turned evaluators) insist that the most reliable and appropriate methods of program evaluation are the methods of social science. In its most extreme formulation, this view proposes that the preferred method of evaluation is "experimental" or "quasi-

experimental." (Campbell and Stanley, 1966; Cook and Campbell, 1975) Only these methods, some believe, are scientific and reliable and will yield valid, objective and impartial results.

A wide range of limitations and difficulties with these methods, however, has been experienced in actual program evaluation settings. Their practical payoff has not matched their early optimistic promise. Consequently, a reconsideration of more reasonable, feasible and satisfactory methods of program evaluation has been occurring.

Scriven (1976), for example, underscores that evaluators "must frequently face the need to do the best we can with non-experimental data." He sketches the *modus operandi* (MO) method of identifying probable cause. This method is familiar in the approaches of the detective, coroner, clinician, historian and anthropologist who employ "causal checklists" and "pattern recognition" to establish probable cause. It also appears to be the method used by the mechanic, doctor, consultant, diagnostician, specialist or trouble-shooter in almost any field. It is likely a major approach used (if only implicitly) by the successful program manager and decision maker. It is seemingly commonplace also as a general method used by the experienced evaluator especially at the point where the powers of formal methods end (and they all do) and probable cause must be inferred from a variety of partial, provisional, complex and incomplete information. Scriven suggests that "the main thrust of efforts toward sophistication [in evaluation] should now turn from the quasi-experimental toward the *modus operandi* approach." (p. 108)

Recently, Alkin, Daillak and White (1979) also expressed a view quite different from those of Suchman and Caro. Wondering why there appeared to be so many "wasted" evaluations, they open their examination of five cases of evaluation utilization this way:

> Why should we even be concerned with this question? [waste] The answer is to be found in the fundamental distinction between "evaluation" and "research." One of the authors of this book, along with others in the field of evaluation, has felt that the clear understanding of the distinction between these two kinds of studies is essential to the development of evaluation theory and ultimately to the practice of evaluation (Alkin, 1973). On the one hand, there are studies designed primarily to add to the body of knowledge (research), on the other, those studies designed primarily to provide information for decision-making (evaluation). And these two functions are separate and distinct. The following typical comment provides a case in point: The study was appropriate even if the results were not utilized since its redeeming feature is its intrinsic value and its contribution to the corpus of knowledge. Such a statement is appropriate as a comment on research but not on evaluation. (pp. 13-14)

Finally, some evaluation experts appear to strike a neutral ground with respect to evaluation *method*. In a recent discussion of the relationship between zero-based budgeting and evaluation, for example, Wholey (1978) emphasizes *systematic measurement* of program performance, but here he does not specify methods:

> In this book we use the term *program evaluation* to mean the systematic measurement of program performance (resource inputs, program activities undertaken, resulting outcomes or impacts), the making of comparisons based on these measurements, and the communication of evaluation findings (measurements and comparisons) for use by policymakers and managers in decisions on government programs. (p. 47)

There are, in short, divergent opinions held by reputable parties about what types of methods (and what canons of proof and evidence) are appropriate to and adequate for program evaluation purposes. These divergencies are explained partly by the existence of different schools of opinion about what constitutes reliable information and knowledge, what constitutes evidence of "cause" and what is the nature of the public agency evaluation problem in the first place.

Table 2 identifies broad groupings of approaches which are available for program evaluation. The groups of methods clearly overlap and are used in varying combination. For example, *all* the methods are dependent on and are used in combination with ordinary intelligent observation and analysis (II). Many employ the *modus operandi* method (IV). The results of all public agency program evaluation must be used through interactive social processes (I). (Lindblom and Cohen, 1979)

Since the problems to which different programs are directed and the conditions under which they operate vary widely from instance to instance, there is no way to prescribe in advance which category

Table 2

Some General Approaches and Methods Used In Program Evaluation

I. Evaluation Through Interactive Social Processes: e.g.,
   - Political Processes of Bargaining and Adjustment;
   - Market Processes;
   - Mixed Processes; and
   - Other Social Processes.

II. Ordinary Intelligent Observation and Analysis.

III. Conventional Methods of Investigation: e.g.,
   - Observation;
   - Fact Gathering;
   - Historical Analysis;
   - Contextual Analysis;
   - Data Synthesis and Analysis;
   - Inferential Reasoning; and
   - "Guesstimating."

IV. "MO" *(Modus Operandi)* Method: e.g.,
   - Use of Implicit or Explicit Causal Checklists and Pattern Recognition by Clinicians, Coroners, Detectives, Troubleshooters, and Others.

V. Systematic Analysis: e.g.,
   - Cost-Benefit Analysis;
   - Cost-Effectiveness Analysis;
   - Systems Analysis; and
   - Policy (Program) Analysis.

VI. Formal Social Science Approaches; e.g.,
   - *Ex Post Facto* Design;
   - Pretest, Posttest Design;
   - Quasi-Experiments;
   - Controlled Experiments; and
   - Others.

VII. Other Methods (Recognized by the Reader).

VIII. Mixed Methods (Combinations of the Above).

of methods would be superior to any other. A general pragmatic rule is *fit the method (tool, approach or process) to the problem*. And not vice-versa. There are at least two steps which can be taken prior to a decision to evaluate a program *formally*: (a) preparation of a brief program evaluation issue (problem) paper, or (b) conduct of an "evaluability assessment." Either of these *pre*-evaluation steps will assist in determining which, if any, evaluation methods and approaches might be most suitable in a specific situation. The steps are discussed in part VI. We turn next to a selected set of basic issues raised by formal program evaluation when applied in practice.

III. Selected Issues Raised by Formal Program Evaluation

Attempts in the last 20 years, of which I am aware, to introduce formalized technical aids to decision making (e.g., program budgeting, policy analysis, planning-programming-budgeting (PPB), management-by-objectives (MBO) and performance monitoring) into the complex, interactive and political environment of a public human service agency have raised a variety of issues, questions and dilemmas. Formal program evaluation is no exception as the literature shows. (Knezo, 1974; Chelimsky, 1977; Patton, 1978) Many questions relate to conceptual, technical, or methodological issues of the *formal* approaches to evaluation. While these issues are important, there is another class of issues which seems basic and persistent; namely, the general problem of "fit" (or misfit) between the assumptions and requirements of traditional formal program evaluation and the processes of political economy which make up the normal environment of public human service agencies. This part discusses a small selected set of these issues: expectations; the so-called criteria and indicators (measures) problems; "causation" as a prescribed focus of one branch of evaluation; the consequences of adapting general program designs to local circumstances; the degree of control which public agencies exert over programs and problems which may be evaluated; and the sense in which program evaluation is "research" and/or "science." The discussion begins with our expectations.

The Role of Expectations

Our expectations heavily color our judgments of the results of what we do, i.e., our evaluations. If expectations are extremely high, we may view modest results with disappointment as shortfall or failure. By contrast, if expectations are very low, it may not take much to satisfy them. The same modest results may now look better—like progress or success. This basic psychological relationship between what we look forward to (expectation) and what we get (results) lies at the base of both individual and collective evaluations.

This phenomenon is akin to the differences in perception by which one may see the same glass of water as either half full or half empty. The same phenomenon occurs when an evaluation result of 25 percent is viewed as either a little or a lot. (Weiss, 1973) In all cases, results depicted by a formal evaluation will be measured against implicit or explicit *expectations* about *anticipated results*.

At the national level, the perceived shortfall or failure of many of the programs of the era of the "Great Society" can be attributed in part to what can be seen in 20-20 hindsight as high, if not unrealistic, expectations about what was both possible and probable. Viewed from a contrasting and optimistic vantage point, Wattenberg (1978) examined a wide array of evidence on changes in individual, social and economic conditions from 1960 to 1976 and found ". . . that behind the harum-scarum headlines a great deal of remarkable progress has occurred in the United States in recent years." (p. xi)

At the level of individual human service programs, it matters greatly whether evaluators, program managers or other influentials who participate in decisions on funding, program design and management expect large results, small ones or none at all. Whether our judgments are that programs work or do not, that they pay off or not, or that they are worthwhile or worthless is in no small part a direct function of our *expectations* about them.

Similarly, when we put our management tools and approaches to tests of worth (for example, MBO, policy analysis, PPB or program evaluation), we judge them better or worse largely as a byproduct of the types and levels of expectations we have about what good they will do in the first place. The history of the "success" or "failure" of program evaluation is itself written against the backdrop of what we hoped, were told, were sold or otherwise came to expect would be its value. It is clear, then, that we can adjust valuations of worthiness of programs not only by changing them or by changing the measuring tools by which they are gauged, but also by adjusting expectations about them.

Values in Evaluation

It is sometimes assumed that formal evaluation will substantially reduce if not eliminate the intrusion of values into decisions about programs. Because it will presumably be based on the "impartial" generation of verified "facts" and conclusions through the application of "reliable" methods, formal evaluation is sometimes viewed as relatively value-free or value-neutral. In practice, however, formal evaluation is, like other modes of research and analysis directed at public policies and programs, not value-free but value-embedded. Here are several of the many ways that values enter, directly and indirectly, into all formal program evaluation processes.

1. Selection of the Program To Be Evaluated

Time, resources, interest and common sense ensure that a public agency does not usually and formally evaluate all of its programs at once. A selection of one or a few from among many is necessary. The motives for an evaluation may be several and usually mixed: to comply with an external mandate from an authorizing or funding source; to verify problems which seem to exist (low morale, drops in productivity, excessive processing times, unusual costs, poor targeting, etc.); to respond to an outside charge about program performance; to inquire into how the program actually works; and so on. Whether there is a single motive or several, a decision to evaluate one program rather than others is an act of *selection*. It focuses attention by subjecting one program to formal scrutiny while sparing others. The resulting attention may change the image, aura, competitive position or other conditions of the subject program compared with others. Both the motives which lead to the evaluation and the act of selection itself are avenues along which values enter early the formal evaluation process. The selection of one program for evaluation over others has been called "a political act."

2. Choice of the Evaluator(s)

For all the reasons that individuals in any craft or profession vary from one to another (skills, experience, competence, motivation, social philosophy, etc.), so do evaluators. The selection process by which evaluators are chosen, whether informal and simple or formal and complex, will, by intention, screen some potential evaluators (and their likely evaluation approaches) in and others out. Since evaluators are not interchangeable, some additional measure of variable value orientation will enter the formal evaluation process at this stage.

3. Negotiations Between the Evaluator(s) and the Client(s)

However focused an initial evaluation proposal, plan or design may be, negotiations between client (sponsor) and evaluator are common and *essential*. In these negotiations, emphases, priorities, measures, approaches, understandings, etc., will be further shaped.

4. Conduct of the Evaluation

Few evaluation studies go precisely according to plan and design. Unpredictable field conditions and barriers, unanticipated staff turnover, data shortfalls, misestimation of logistical and time requirements, changes in the sponsor's mind and so on are the common challenges to evaluation management. Substituting proxy measures for intended ones, modifying a planned sampling

method, trimming and redesign will all contribute to further shifts in scope, emphasis and approach. Though some accepted guides and technical adjustments to field study exist as standard procedure to deal with so-called "threats to internal and external validity," the path of field work is rarely smooth or according to plan. It is normally bumpy and strewn with compromises.

5. Analysis of Results

It is a common misconception that "data speak for themselves." Yet, like two witnesses who report the same accident in different ways, two evaluators may interpret the same findings and data in dissimilar ways. Members of the *same* evaluation team often come to different interpretations of the same data. They may thrash out differences and compromise for the sake of a show of consensus and unity in the final report. Some evaluation experts suggest that this practice robs the user of legitimate alternative (divergent) interpretations and possible insights which may prove valuable if not decisive. They urge the open submission of minority reports as a routine part of the presentation of study findings. Re-analysis by outside analysts of the data from a completed evaluation may (and often does) turn up new interpretations and conclusions.

6. Inferences From Findings to Recommendations

Discovering what *is* tells us little or nothing about what *ought* to be. Whether the findings of an evaluation study are descriptive or explanatory, there are no ready-made rules for moving from descriptive statements of "fact" to prescriptive statements about what ought to be done in the future. Since program evaluation is carried out in a value-diverse environment of competing claims for public and social resources, the formal evaluator must invoke assumptions, chains of reasoning, supplementary knowledge, theory and social philosophy to move from findings to recommendations about program change. The strict researcher may be inhibited by professional norms from getting "too far beyond the data." Yet public agency users are often interested in moving well beyond the data to guidance about "What should we do now (next)?" Depending on its length, the inferential leap from so-called fact or finding to prescriptive action may traverse a lot of value territory.

7. Use of Study Findings in Policy Debates

If and when evaluation study findings are invoked in policy discussions, the net of interpreters is enlarged, usually well beyond the original client and evaluator(s). New actors usually bring somewhat different perspectives, assumptions, chains of reasoning, experience-based knowledge, incentives and social philosophy to the interpretation of study results. Since no study ever covers the waterfront of a program or presents findings and conclusions with equal clarity, evidence and certitude, the terrain of possible variable interpretation is substantially enlarged at this step in the program evaluation process.

There are, in short, not a few but many complicated, blatant and subtle ways that values enter even the most technically and managerially scrupulous program evaluation process. This is not a problem unique to formal program evaluation. Steps can be taken, for example, to keep major value shifts and drifts as explicit as possible, avoid gross instances of willfull bias and subject resulting work to scrutiny from many points of view. Despite these efforts, program evaluation will remain value-embedded rather than value-free or value-neutral. [Note: For one detailed attempt to cope self-consciously with some of the value issues in practice, see the history of the attempt by the National Institute of Education to evaluate the ESEA Title I program (compensatory education) for the Congress (Pincus, 1980).]

Thus far we have brushed past a major step in formal evaluation where value issues and technical issues are frontally or subtly joined. This occurs in the inevitable selection of the explicit criteria, indicators and measures in terms of which program performance and impact will be formally examined.

The Criteria and Indicators Problem

Traditional formal evaluation usually requires explicit measurement. Every attempt to gauge the effects or impacts of a program *formally* requires the selection and specification of some *criteria* (e.g., income, achievement, health status, etc.) in terms of which the output, outcomes, impacts, effects or results are to be examined. But evaluation methodology is silent about what is to be measured. For example, how should the effects be gauged of an educational program ostensibly intended to improve student learning? What criteria should be invoked? The answer to these questions turns partly on your point of view. When economists attempt to answer this question, they often turn naturally to those criteria about which they know most—economic criteria. They may try to estimate the change in future earnings attributable to, say, a program intended to reduce the high school dropout rate. By contrast, the bulk of the attention of "cognitive" educational evaluators has historically been paid to the effects of programs and practices on student school performance measured largely in terms of grades and achievement test scores. By further contrast, a general psychologist might look at the effects of the same program in terms of its influence on "noncognitive" factors such as a student's sense of self-esteem, sense of control of his/her environment, or attitudes toward risk-taking and uncertainty. Others may look to effects of a program or practice on personality "traits" such as assertiveness, persistence, sense of responsibility and self-control. Still others may look at program consequences for general social skills required for interpersonal adjustment, coping with change and stress management.

Once criteria have been chosen, for whatever reasons, specific *measures* or *indicators* must be selected to reflect these criteria. If student "achievement" is to be measured, for example, which of the many existing achievement tests should be employed? The formal evaluator may focus on the important issues of technical validity and reliability of a given test. But selecting tests has not only technical aspects but value dimensions as well.

Although educational research has been going on for decades, there is no consensus about *what* should be measured or by which tests. So-called standardized achievement tests are deeply rooted in the educational system. But some critics of standardized tests raise basic questions about the extent to which they actually measure learning ability and accomplishment rather than the acquisition of knowledge about the content of the dominant culture. Most experts agree that culture-*free* tests are impossible (since education is itself a part of culture). Some do, however, propose tests that appear in their view more culture-*fair*. Debate does not then stop, but normally shifts to what is "fair" and what is "foul." The many debates about tests (their assumptions, their methodology and the social consequences of their use) reflect in turn larger issues about (a) what purposes education does or should serve in the society at large and (b) what influences educational philosophies and practices have on the attitudes, values, skills, competencies and futures of individuals in the educational system. Many and diverse economic, social, cultural and ideological views of education have come vividly to the fore in policy and social debate in the last 20 years. They have been heated and virulent.

The point is clear: there are *alternative* and *competing* criteria and measures in terms of which the impact on students of an educational program can be measured. Reflecting on the more general role of education in society, Jerome Bruner, noted education expert, underscores the culturally embedded and value-laden nature of the study of child rearing and human development. His remarks amplify and set into a more encompassing framework the comments just made about the education criteria and measurement problems:

> I would urge that in the nurturing of the young, a society is required to make a continual series of decisions about its norms. Child rearing is neither a private activity nor is it "factual" nor dispassionate. Since human development is as much determined from the outside in as from the inside out, its guidance is as much a prerogative of the culture, as it is a reflection of the intrinsic growth of the nervous system. . . . It is in consequence of this position that the study of human development is so implicitly guided by policy needs: how to raise or even define an intelligent human being, how to assure the growth of a proper moral judgement or an adequately evolved logical capability, how to increase

> independence or loyalty or tenderness, how to prevent alienation and anonymity in a technological order or to maintain identity in the face of urban mobility. . . . So long as the culture arranges the conditions under which growth is supposed best to occur, the study of human development must in some inevitable way be a normative science, a policy science like economics. The shape of the sciences of human development, either in the past or in the future, will to some considerable extent be a result of subtle (and sometimes not so subtle) forces imposed upon it by the culture in which these sciences exist. (Bruner, with the author's permission.)
>
> [Note: Neither the author of this monograph nor the author of this quote was able to relocate the exact source. Similar ideas can be found in Bruner (1971).]

In sum, the evaluation measurement problem requires the selection of (decisions about) measures and indicators which have *analytic* utility; that is, which presumably enhance understanding of the phenomena they are designed to measure. The problem also raises *value* questions about which aspects of individual and social development on which programs impact are worth measuring in the first place. In program and policy studies, analysis and evaluation, the so-called criterion problem and its allied measurement problem are at their heart also value problems. In practice the choice of measures may be made in some inductive and pragmatic way; the analytic and value questions are "decided" simultaneously.

The existence of stated program goals and objectives may limit the range of choice among criteria and measures. But program goals and objectives are often aspirational and the result of political compromise. As a result, they tend to be general, abstract, multiple, often conflicting and evanescent. Those who initiate and conduct evaluations are ordinarily forced to choose, if not invent, alternative criteria and measures for evaluation purposes. While there are no universal operational guides to these choices, here are five general rules of thumb which appear useful.

1. Employ measures which are of expressed interest to the evaluation sponsor(s) or expected users.

2. Employ multiple measures, when feasible, rather than just one.

3. Acknowledge the value implications of selecting criteria and measures and make value decisions openly and consciously.

4. Consciously select measures from more than one value set. In the education example, this may imply choosing not only from among "cognitive" measures of performance (such as grades and achievement scores) but also from among "non-cognitive" measures as well (such as measures of self-esteem, sense of "internal-external" control, etc.).

   It is worth noting that while cognitive and non-cognitive measures refer to different sets of factors which may be associated with learning progress, their mutual interactions and relative contributions to learning are still basic open questions.

5. Keep intangibles in the foreground and not the background of the analysis. Many aspects of programs may be judged important and yet not be susceptible to measurement. *Measurability* is in no way an index of *importance*. Tellingly, Campbell, a major exponent of experimental and quasi-experimental evaluation research, carries the point further:

   > Too often quantitative social scientists, under the influence of missionaries from logical positivism, presume that in true science, quantitative knowing replaces qualitative, common-sense knowing. The situation is in fact quite different. Rather science depends on qualitative, common-sense knowing even though at best it goes beyond it. Science in the end contradicts some items of common sense, but it only does so by trusting the great bulk of the rest of common-sense knowledge. Such revision of common sense by science is akin to the revision of common sense by common sense which, paradoxically, can only be done by trusting more common sense. (Campbell, 1979, p. 70)

Evaluation and "Causation"

The branch of evaluation which flows directly from the formal research components of some social sciences carries with it an aspiration to establish the "causes" of program effects. While the search for causes of events appears to be a central concern of some social science research, it seems beyond the power and ken of applied program evaluation. The reasons are several. First, establishing "causation" in a meaningful way is a complex problem which is itself a subject of study by methodologists, epistemologists, and philosophers of science. Second, the social sciences reportedly have great difficulty establishing reliable and verified causal relationships even in the case of the more constrained problems studied in "the laboratory." (Campbell, 1979; Almond and Genco, 1977) Establishing formal "causes" in the open, contingent, evolving and highly interactive world in which real programs operate is a much more strenuous and complex task. Third, well short of establishing the "cause-effect" of specific individual programs, there are major and unsettled questions of causation associated with nearly all the basic human, social, economic and cultural problems that are the targets of human services programs in the first place: incentives for human learning, mental illness, alcoholism, delinquency, work incentives, ill health and the like.

For example, some health programs encourage and/or pay for visits to the doctor. While doctors play an important role in circumstances such as acute illness and medical emergency, medical services appear to account overall for substantially less than 10 percent of the variation in the health status of the population. Acknowledging the important role of genetics (Luria, 1973), most experts tend to agree with Knowles (1977) that "The health of human beings is determined by their behavior, their food, and the nature of their environment." (p.57) We do very much more to determine our own good or ill health than medical services do for us. According to Knowles, "Prevention of disease means forsaking the bad habits which many people enjoy — overeating, too much drinking, taking pills, staying up at night, engaging in promiscuous sex, driving too fast, and smoking cigarettes. . . ." (p. 59) We seem intent on making ourselves sick.

In a similar vein, Lewis Thomas, president of the Memorial Sloan Kettering Cancer Center in New York City, reviewed the "science and technology of medicine" and concluded that despite progress:

> We are left with approximately the same roster of common major diseases which confronted the country in 1950, and although we have accumulated a formidable body of information about some of them in the intervening time, the accumulation is not yet sufficient to permit either the prevention or the outright cure of any of them. (In Knowles, p. 37)

Health promotion and disease prevention appear to lie not dominantly in medicine but much more heavily in our relationship with our environment and in the things we do to and for ourselves. (Eckholm, 1977; Sobel, 1979; Dubos and Escande, 1979)

A fourth factor which inhibits the evaluator from reliably attributing "causes" directly to program effects is the multiplicity of other programs which may be operating in the environment of the program under evaluation. The interaction effects among the several programs may make separating the effects of any single one of them impossible. Gorham, president of the Urban Institute, and Glazer, an associate, point to this problem in an examination of the "urban predicament." (1976) Based on their review of poverty programs of the late sixties and early seventies, they concluded that "some of the changes that reduced poverty from 19.3 percent in 1964 to 12.8 percent in 1968 could be traced in part to the poverty program." But they also pointed to a problem endemic to imputing "cause" to a specific program intervention which operates in a complex environment:

> The overall evaluation of poverty programs and model cities will probably always be in dispute. Reliable evaluation is hindered by the fact that the poverty programs were only a few of a large number of factors affecting income and "opportunity." Of the other forces that were increasing income, perhaps the most important was the very high economic growth the United States was experiencing during the late 1960's. Equal opportunity was given a great boost by the passage of new civil rights legislation in 1964 which banned discrimination in employment. Sorting out the contribution of each of these factors is, at least for the present, beyond our most perceptive evaluators. (p. 11)

A fifth and final factor (mentioned here) which muddies "causal" waters is the extent to which any given program design takes on highly diversified and variable configurations from setting to setting as general elements of design are fitted to local circumstances. The phenomenon of local adaptations deserves further comment.

Local Program Adaptations

In terms of the *specificity* of their designs, public human service programs range from the general (block-type and formula grant) to the highly specific (such as programs for remedial reading). The design of many others falls somewhere in between. The programs reflect a basic, though often general, structure, may include prescribed elements (and other design features), and are often directed to a target problem such as drug abuse and/or a target group such as elementary school children. Examples include community mental health centers, neighborhood health centers, compensatory education programs and a wide range of other State and locally operated human service programs.

These "model" designs are often based on the assumption that they can be implemented (much as they are) in wide-ranging local contexts. But few if any programs fill in the fine details of structure, scope, level, intensity, pattern of management, program configuration, staffing arrangements and the like — all of which and *more* are required to turn skeletal designs into viable, operating service programs. They do not, because they cannot, supply a detailed operationalized recipe for effective implementation.

Program designs must be fitted, tailored and adapted to suit specific, concrete local circumstances or they will fail. As a consequence, formal evaluators who have looked closely have found, sometimes to their dismay, that the variability from site to site of program operations is large if not enormous. In a recent attempt to identify the types of and preconditions for "coordinated planning" between health and mental health planning in nine States, for example, evaluators reported that the diversity and number of conditions and barriers was "almost overwhelming." (Hagedorn, 1980) Similarly, formal evaluation approaches which employ statistical measures to summarize across large numbers of projects fail to capture the important individual character and variation in local program adaptations. Instead, they often successfully mask, rather than disclose, what seem to be among the basic determinants of program performance.

McLaughlin (1980), for example, appears to have come to this conclusion after participating in a Rand study of several education programs: ESEA Innovative Projects, ESEA Bilingual Projects, Vocational Education Exemplary Projects and the Right-To-Read Program. The Rand team spent 2 years examining local projects under the four programs and 2 additional years following those projects under the two largest. Team member McLaughlin concluded that conventional formal methods of evaluation do not work. She argues that many evaluation approaches merely assume a "black box" between project or program inputs and their outputs or effects. She concluded: "The contents of the black box, it turns out, matter more to project outcomes than do other factors that evaluators attempt to calibrate and assess." (p. 42) She identified "a number of factors that are generally ignored in special project evaluations, but that are required for a valid evaluation design." They are factors not of program design but of the interaction of elements of a local context or program setting:

1. *Institutional support and receptivity* which included administrators' attitudes and support of the project and a broad-based implementation strategy which involved all significant actors and resulted in staff commitment;
2. The baseline *capacity and expertise* which local staffs possess at the start of a project and which may vary widely; and

3. Available *local implementation choices* about how the project will be put into practice. Successful "choices," in these cases, seemed to include training tied directly to the concrete problems, expressed needs and suggestions of participants; concentration in a few sites (rather than a shotgun approach) and a "critical mass" of supporting participants; locally prepared rather than imported project materials; and routine staff involvement in decision making about project approaches and materials. (pp. 43-44)

The preconditions of participation, timely feedback (which allows error correction), involvement and commitment made implementation "heuristic—a process of learning and adjusting, rather than a process of installation." (p. 44) Other local conditions and events unrelated to project design (for example, institutional climate, leadership style, cutbacks, teacher strikes) make up the *natural setting* within which programs and projects are implemented. They heavily influence whether a project makes a difference. McLaughlin concludes:

> Yet these local factors are seldom in project evaluation models. . . . A special project cannot be validly assessed in isolation from its system context. (p. 45)

A direct corollary of the adaptations which must be made to make programs work satisfactorily is their changing, evolutionary and developmental nature. Programs have natural histories. Fledgling programs and projects operate differently from mature ones, though change may be regressive as well as progressive. A continuing series of adaptations to changing program and project circumstances leads to a regular succession of different program operations and configurations. [Note: See White (1977) for an excellent brief account of the impact of changing context on an initial attempt—which apparently failed—to install a performance (evaluation) monitoring system in a large urban school system.] Impact, outcome and output-oriented studies usually miss the contextual and operational dynamics of programs and thus the circumstantial and adaptational features which help explain program performance.

Limits on Public Agency Control

It is commonly assumed in much of the writing on evaluation that if human service programs are formally evaluated and found to fall short of performance standards or expected results, public agencies can correct them. In practice, a large share of the problems to which public programs are directed are complex and varied in cause. The factors which give rise to many of them lie well beyond the reach of government in general and beyond the influence and control of specialized agencies.

Similarly, public agencies are not autonomous agents free to change their programs and policies at will or on a moment's notice. They operate instead within the familiar environment of social, economic and political pressure and commitments which arise not merely from citizen wishes but also forcefully from legislatures and councils, service providers, professional associations, commercial, financial and industrial beneficiaries, other levels of government, other agencies, ambient electoral, party and coalition politics and so on.

Some of the pressures which appear to be external are heavily articulated through the internal management and politics of the agency which itself consists of additional sets of "stakeholders." The interplay of internal and external pressures and the operating commitments of an agency substantially circumscribe freedom to act either directly or decisively. Despite the size of a budget or the illusion of command and power, agency officials often have influence over only a small (though sometimes important) share of an agency's resources or of its overall operations. As more social functions have been performed by public agencies, the number and hold of "stakeholders" on agency resource allocation and management decisions appear to have increased and reduced the discretion of agency leadership.
[Note: We thank Steven J. Brams of the Department of Politics, New York University, for emphasizing the notion of "stakeholder."]

In short, individual public agencies have significant but circumscribed control over their own internal operations and little, if any, over many of the problems they attempt to ameliorate or contain through their services. As a consequence, global evaluation studies directed at society at large or at the "root" causes of generic individual and social problems will probably yield less than we collectively now know and not much of practical value to an operating agency. Limits on agency control help explain why some past evaluation efforts have had little impact. Many program problems lie beyond the reach of public agencies. As a consequence, program issues or problems should be selected for evaluation partly in terms of the extent to which the sponsoring agency can influence the factors associated with ameliorations or remedies. Public agencies neither control everything nor control nothing. Evaluation studies should be targeted to areas and subjects about which agencies can reasonably be expected to have some say.

Is Program Evaluation Research and/or Science?

These questions may provoke word quibbles. But they also reflect additional evaluation issues which have taken on philosophical, territorial, economic and political overtones. In its root meaning, "re-search" (from Old French *recerche*) means "to seek out, to search again." The *American Heritage Dictionary* (Morris, 1976) indicates that the word may refer to "scholarly or scientific investigation or inquiry," or it may mean "to study thoroughly." It is in the more open and general meaning of *study* that program evaluation can be usefully understood.

In a narrowly restricted sense, the view of evaluation as "scientific research" naturally raises the corollary question, "Is evaluation science?" In terms of the actual practice and performance of program evaluation, even in its most exemplary form, few experienced evaluators would argue with a flat "no." Though it may employ technical methods, formal program evaluation, like its cousin policy analysis, is art and craft and not science. (Wildavsky, 1979)

A related and underlying question is, To what extent is "social science" science? Pursuing this question might carry us afield, but it raises deep questions: What are the bases and alternative paths to human understanding and knowledge? What constitute reasonable and workable canons of proof and evidence in social science on the one hand and in political and bureaucratic decision making on the other? What kinds of proof and evidence about what kinds of subjects matter in the political economy of public agencies? These are more than idle questions. (See Lindblom and Cohen, 1979; Almond and Genco, 1977; Campbell, 1979; Sharpe, 1976; Rein, 1976; Thorson, 1970.)

Summary and Conclusions

The discussion to this point deserves a brief summary and the restatement of some conclusions.

Stimulated by Federal requirements for and financing of formal evaluation as a precondition for financial support of services, evaluation activities have grown by some accounts into a well-financed industry. Though formal program evaluation has been mandated widely by the Federal Government (and more recently by some State governments), ambiguity and uncertainty persist over what means should be employed to generate acceptable evidence of program worth and value. Early on, proponents of evaluation as formal *research* came to dominate the literature on evaluation which contains many authoritative prescriptions. For the most part they urged the use of traditional social science research methods. Proponents of these methods often assert that they will generate impartial, objective, verified and reliable information on the impacts and effects of programs and will help identify their "causes." This information, it has been assumed, would, by sheer force of its authority and weight, lead to program improvements by "rationalizing decision making."

Authoritative prescriptions for evaluation research have not led to notable successes, but rather to disappointing and meager results. Casual evidence, a growing body of testimony from experienced evaluators and evaluation sponsors, and the results of several studies suggest that the actual performance of formal program evaluation has fallen far short of its early promises. Excessive government mandates coupled with dubious advice about preferred "scientific research" methods for evaluation seem to have led to frustration, some waste, very modest results and growing calls for fundamental reforms in the philosophy and practice of program evaluation.

As it turns out, the basic activity of evaluation — ascertaining or fixing the worth or value of something — is a commonplace, everyday human activity. Difficulty ensues when concern with evaluation is shifted from the level of the individual to the level of collective social judgment. At a collective social level, there exists a wide variety of political, economic and social mechanisms and processes through which judgments are regularly expressed about the worth and value of social programs. Although proponents of formal rationalistic approaches to evaluation find these mechanisms faulty and wanting, they remain among the most dominant, available and widely used vehicles by which collective social judgments are expressed about the use of resources in public programs. And they are the mechanisms and processes through which the results of formal evaluation must be used, if they are to be used at all.

Many textbook models of formal evaluation appear to derive from an idealized, technical and linear (sequential) style of thinking and problem solving which does not "fit" well the environment of social and political interaction and adaptation in which all public programs operate. Although many social scientist-evaluators claim superiority for their preferred methods, a listing of some of the available alternative approaches, methods and mechanisms for evaluation includes existing political and bureaucratic processes, the exercise of ordinary intelligent observation and analysis, the use of conventional and widely available methods of study and investigation (including the widespread use of implicit or explicit causal checklists and pattern recognition), the use of a variety of modes of systems and policy analysis, and many *combinations* of interactive and analytical methods of social problem solving.

Attempts to apply formal research methods of evaluation in practice raise many basic issues. A brief discussion of some of them revealed the following:

1. Above and beyond the *methods* of evaluation which might be employed, our *expectations* heavily color and influence our judgments of the results of what we do, including evaluative judgments about social programs.
2. Despite myth and rhetoric, formal evaluation is not value-free or value-neutral. It is, instead, value-influenced and value-embedded. The many avenues through which values enter the practical processes of formal evaluation include (but are not limited to) the selection of a program for evaluation in the first place, the choice of evaluators (and their preferred approaches), negotiations between evaluator and sponsor, compromises and adjustments required by field work, inferences required to move from findings to recommendations for future action and variable interpretations of the same study findings ordinarily made by a variety of actors involved in agency program decision making.
3. The selection of criteria and indicators (measures) in terms of which program performance evaluation might be made involves a set of *both* technical and value judgments which are intertwined and inseparable. Selecting criteria and indicators is not a mere technical problem but at its heart also a value problem.
4. The branch of formal evaluation which derives from social science research traditions aspires to establish the "causes" of program effects. Yet an inquiry into our knowledge about a wide array of individual and social problems at which public programs are directed suggests that knowledge is normally partial, provisional and often conflicting. In addition, many other programs may operate in the immediate environment of the one under evaluation and separating effects in any reliable way may be extremely difficult if not

impossible; "causation" is an active subject of study and debate by methodologists, scholars and philosophers; and under the best of "laboratory" conditions the social scientist appears to have great difficulty establishing reliable and verified "causal" relationships even in the case of carefully selected and constrained research problems.

5. Though the philosophy of formal evaluation often assumes that a program *design* will be implemented in recognizable form amenable to easy detection and study, a variety of evidence suggests that the actual configurations and features of operating programs are the result of varied and complex adaptations to specific local circumstances and conditions. Many conventional methods of formal evaluation, especially summary statistical methods, appear to miss or mask the very factors which appear to contribute to effective program operations. These insights have led to a growing number of proposals for major reform of evaluation theory and practice.

6. Much writing on program evaluation appears to assume that the program defects and problems uncovered by formal evaluation can be corrected by the public agencies which finance and operate them. In practice, however, individual public agencies have only circumscribed (though often significant) control over their internal operations and little, if any, over many of the problems which they attempt to contain or remedy through their services.

7. Assertions to the contrary aside, a dozen years of experience with actual practice suggests that program evaluation is art and craft and not science. An examination of some of the prevalent problems of "fit" between formal research methods and program evaluation carried out in actual public agency settings raises provocative questions about the extent to which the social sciences are "science."

Next is a brief examination of some of the evidence about the difference that formal evaluation appears to make in practice followed by an identification of a few of the many proposals o reform evaluation. The reader not interested in the details of studies may want to page ahead to the **Intergovernmental Lessons** recited at the end of part IV or turn directly to the guidance for the practitioner presented in part VI.

IV. What Difference Does Program Evaluation Make in Practice?

If program evaluation is valuable and worthwhile to a public agency and to its decision makers, its value and worth should be shown through the use to which evaluation results are put and through the impact which evaluation has on the opinions, attitudes, decisions and actions of policy makers.

Unfortunately, though the literature on evaluation is large, and the claims for its value numerous, there are surprisingly few documented studies of its impact. To find them, a broad-based key-word search of several large abstract and information services was conducted. It yielded about 620 individual abstracts.

Project SHARE	130
HEW Evaluation Documentation Center	280
National Criminal Justice Reference Service	175
Dialog	35
HUD USER	0
Total Abstracts	**620**

We screened these abstracts for relevant sources, received useful suggestions from interviewees and colleagues, scanned issues of *Evaluation* magazine from 1972 to 1979, perused several evaluation journals and reviews, and drew on our own library. In all, we examined in hard copy over 150 evaluation sources including manuals, case studies, articles, books and papers. We have referenced only a tiny fraction of this material.

As a result of this partial but extensive search, we found no body of systematic or social scientific studies which yield a "valid" and "verified" picture of the utility, uses, outcomes, impacts and side effects of program evaluation in a public agency context. This may seem surprising in light of over a dozen years of experience with evaluation, the expenditure of billions of dollars on evaluation studies and the insistent demands by proponents of evaluation that expenditures of public funds should be put to the formalized tests of evaluation to assess their impact, establish their worth and improve their relevance and utility. How would program evaluation hold up under the scrutiny and demands for evidence of worth and value which program evaluation is intended to bring to bear on public programs generally? There appear to be no definitive answers to this question. In addition to the testimony of experienced evaluators cited earlier, however, there is a small body of partial and fragmented evidence which contains some clues.

Table 3 lists the major sources of evidence identified in the literature search. The list is followed by a brief summary of the findings of each study and includes occasional comments on limitations of the study or on its apparent significance. The sources are presented one by one roughly in the chronological order in which they appeared. Since the sources are few, no overall summary is provided.

Table 3

Sources of Evidence on the Impact of Program Evaluation

Municipal Management and Budget Methods: An Evaluation of Policy Related Research, Final Report. Volume 1: Summary and Synthesis (Kimmel, Dougan and Hall, December 1974).

Program Evaluation Within California State Agencies: An Assessment (Conner, Rosener and Weeks, May 1976).

"Symposium on 'The Research Utilization Quandary' " (Weiss, Spring 1976).

"Factors Associated With Knowledge Use Among Federal Executives" (Caplan, Spring 1976).

Assessment of State and Local Government Evaluation Practices in Human Services (Baumheier et al., February 1977).

Interim Analysis of 200 Evaluations of Criminal Justice (Larson et al., May 1979).

"Lessons Learned From Federally Mandated Program Evaluation for Community Mental Health Centers: Framework for a New Policy" (Flaherty and Windle, May 1980).

Utilization-Focused Evaluation (Patton, 1978).

Summaries of Selected Studies

Municipal Management and Budget Methods: An Evaluation of Policy Related Research, Final Report. Volume I: Summary and Synthesis. Volume II: Literature Reviews. (Kimmel, Wayne A.; Dougan, William R.; and Hall, John R. Washington, D.C.: The Urban Institute, 1974.)

A team of two analysts and two consultants at the Urban Institute under the direction of the author conducted an extensive literature search for "research on the impact, utility, and effectiveness" of six management and budget methods which might be employed by local governments. The study was one of 19 funded by the National Science Foundation to screen what they described as a "large body of research on municipal systems, operations, and services" created over the last quarter of a century. Each study was to locate, evaluate for internal and external validity, and synthesize for wide dissemination the findings in each area.

The results of the literature search and review of *program evaluation* are summarized this way:

> A search of the literature revealed few empirical studies of the utility, impact or effectiveness of performing program evaluation. One attempt to analyze the impacts of several evaluations was made by Wholey (1973). Ten evaluations were examined in an effort to relate the type of evaluation performed to the influence exerted on budget levels, service delivery and internal government processes. Assessed by the author, four of the ten evaluations appeared to have some impact on budget levels, five on service delivery and six on internal government processes.

In two dissertations, McLaughlin (1973) and Pearson (1973) examined evaluations of programs funded by the U.S. Elementary and Secondary Education Act. The authors judged evaluation to have succeeded or failed in terms of whether it exerted an impact on the management of programs or on later policy proposals. Neither study judged evaluation to have succeeded.

The General Accounting Office (U.S.G.A.O., 1971) reviewed twenty-four evaluations (fourteen completed, ten ongoing) performed for the U.S. Office of Education and found that, in the opinion of OE officials, five of the fourteen completed studies were "of limited use" while the results of the other nine were "adequate and useful."

A study by Eaton (1962) reported an unwillingness among professionals in two bureaucracies (California Department of Corrections and Western V.A. offices) to disseminate evaluative findings that might be considered discouraging or might reflect unfavorably on their organizations. Although there appear to be some weaknesses in the design of this study, its findings relate to the potential utility and effectiveness of evaluation. Evaluation cannot exert an impact on program and policy decisions if findings are suppressed by the organizations for which the evalutations are peformed.

In a related vein, a dissertation by Nielsen (1972) took for granted that most evaluations had exerted no impact on program directors and attempted to discover why. He found that "non-use" of evaluative findings seemed to be due in part to mismatches between the information generated by evaluation and information "needed" by program managers. This explanation was offered as a companion to the frequent observation that program managers are threatened by and hostile toward evaluation.

There appears to be, in short, a very limited body of evidence *from research and formal study* on the utility, impact and effectiveness of conducting program evaluation. (pp. 37-38)

The discussion of program evaluation in the report ends with this concluding note:

The evolution of the literature on program evaluation from the mid-1960's to the present [1974] appears to reflect a disappointment in the capacity of formal evaluation to revolutionize public decision processes. This may stem from a combination of an early "overselling" of evaluation's potential and growing awareness of its sometimes severe limitations.

A single rule-of-thumb for potential users of evaluation might be that the probable "benefits" of an evaluation ought to exceed its "costs" however these are determined. Programs to be evaluated should be of sufficient budgetary importance that it is worth the cost of formally evaluating them. Furthermore, there ought to be a reasonable prospect of a future decision to which evaluation findings can be brought to bear at the appropriate time. Local managers should remember that program evaluation, like anything else, is not infinitely valuable. It may serve a useful purpose in the overall management processes of local government, but only within its technical and political constraints. Not all government programs can or should be evaluated. (p. 47)

Program Evaluation Within California State Agencies: An Assessment. (Conner, Ross F.; Rosener, Judy B.; and Weeks, Edward C. Irvine, Calif: Public Policy Research Organization, University of California, May 1976.)

In this small scale survey, 17 departments, boards and commissions were selected from among 79 in California based on a "judgment" about their high impact on social problems and/or on citizens and on information that they were in fact carrying on some kind of effectiveness measurement activities. Two-member teams, using a 25-question guide, interviewed 16 agency evaluators. They also read and analyzed 43 evaluation reports.

The bulk of the 36-page report of this survey is devoted to a presentation of the answers of evaluators and to a set of recommendations for improving the organization, centralization, visibility, staffing, training and coordination of evaluation. Of interest here are a few findings on the perceived utilization of evaluation results.

1. Six evaluators said results were "very well utilized," nine said "somewhat used" and one said "very little used." The authors summarize: "Current evaluation results, then, are used but not to any great extent." (The study gives no indication of what was meant by "use" or "utilization.")
2. The most frequent reasons given by evaluators for the limited use of evaluation results were: "no incentive" (4); "low reliability" (4); and "results not timely" (3).

3. As to potential benefits of evaluation in the near future, eight evaluators responded a very good likelihood, six said a good likelihood and two said the likelihood was poor.
4. The evaluators believed that other State officials (program managers, department directors, agency secretaries, Governors' officers and legislators) viewed evaluation as it was "currently practiced" as "somewhat useful." They were more optimistic about evaluation as it might be "ideally conducted."
5. Thirteen evaluators viewed the department director as the prime beneficiary of program evaluation and 11 included the program manager. Most did not view the legislature, agency secretary or public and program clients as beneficiaries. (pp. 12-14)
6. Few departments had conducted formal program evaluations; most of their effectiveness measurement apparently took the form of "status monitoring." The authors thought more formal evaluation would be undertaken in the future. (p. 20)

Comments:

The authors clearly favored increasing formal program evaluation activities, especially through the use of control and comparison group studies. The report contains many recommendations to centralize, coordinate and enlarge the evaluation function. Yet there is no show of or reference to evidence that evaluation will pay off to an agency beyond assertion, the reported beliefs of evaluators and the inclusion of a one-page description of a California study of "some additional factors influencing the effectiveness of warning letters" in reducing traffic accidents and convictions. Importantly, the study does not illustrate or describe what was perceived to constitute "use" or "utilization."

"Symposium on 'The Research Utilization Quandary.' " (Weiss, Carol H., ed. *Policy Studies Journal*, Spring 1976.)

Though research utilization is an area of broader concern than the utilization of program evaluation, there are many issues of overlap and common concern. Thus, the reflections of Weiss on the symposium are relevant.

Through a presentation of six papers, Weiss attempted to bring some government officials into a discussion which had been dominated largely by academic social scientists. She also tried to assemble empirical cases to offset the fact that much earlier discussion had been "impressionistic and speculative." Of the six cases, two relate to evaluation studies. Caplan's is discussed in the next section.

The other five included a survey of social scientists (Useem), a discussion of use (Janet Weiss), a view of research use in the State Department (Uliassi), a case study of evaluation of several housing projects (Banks and Clark), a case study of evaluation of an experimental education project (McGowan), a case history of the role of research in mental hospital deinstitutionalization (Swan) and an account of the development and use of research in regional waste water management development. (Conway et al.)

Weiss assesses the implications of the cases this way:

> And what is the verdict from the six case studies about the usefulness and use of social research? Two of the papers are unflinchingly optimistic (Conway et al. and Null), although the evidence in each case is modest. Two find some positive effects of social research, although not necessarily what either the researchers or the sponsors intended (McGowan and Uliassi). Two deal with what might be called utilization fiascos (Banks and Clark, and Swan, but Swan sees hope for the future given the lessons learned). . . .
>
> The theme that emerges from the total set of papers is that the use of research in governmental decision-making is a complex and difficult matter. . . . There is work to be done to clarify the ways in which social research can contribute more effectively to policy. . . . (pp.222-223)

Of the six cases presented in the symposium, the survey by Caplan deserves further attention.

"Factors Associated With Knowledge Use Among Federal Executives." (Caplan, Nathan. *Policy Studies Journal*, Spring 1976.)

This is a summary presentation of a study reported more extensively elsewhere (Caplan et al. 1975). Caplan and associates conducted 204 interviews with officials in Federal executive departments, major agencies and commissions. Interviews were focused on "the use of *empirically* based social science knowledge." The study identified 575 "self-reported instances of social science knowledge use that impacted on policy decisions." Caplan cautions the reader that the findings may be oversimplified.

1. He concluded that utilization (undefined) is most likely to occur when the decision-making orientation of the policy maker is characterized by a reasoned appreciation of the "scientific" and "extra-scientific" aspects of the policy issue. The "scientific" aspect refers to the "internal logic" of the policy issue (a diagnosis of the problem). The "extra-scientific" aspect refers to the "external logic" of the policy issue (the political, value-based, ideological, administrative and economic considerations involved). Caplan grouped officials into three "orientations":

- Twenty percent expressed a *clinical orientation*. They first gather the best available information to diagnose the internal logic of the problem. Then they gather information bearing on the "external" logic of the problem and "finally weigh and reconcile the conflicting dictates of the information."
- Another 30 percent of the interviewees were classified as having the *academic orientation*, those who "are often experts in their field and prefer to devote their major attention to the internal logic of the policy issue. They are much less willing, however, to cope with the external realities that confound policymaking." They apparently use social science information in "moderate amounts" and "in routine ways to formulate and evaluate policies largely on the basis of scientifically derived information."
- A third group, comprising another 20 percent of the interviewees, had the *advocacy orientation*, those "at home in the world of social, political, and economic realities." They reportedly make "limited" use of social science information and "largely to rationalize a decision made on other grounds." (p. 230)
- The orientation of the remaining 30 percent is not provided in this particular reporting of this study.

2. Caplan reports that "the most frequent users of social science research" have a "social perspective — a sensitivity to contemporary social events and a desire for social reform." He comments:

> It is evident to a large extent that many respondents fail to distinguish between objective social science information from subjective social sensitivity. Thus most of the examples which they offered to illustrate knowledge applications really involved the application of organized common sense and social sensitivity, which as a mixture, might be called a "social perspective." (p. 231)

Caplan reports that these officials applied a "value-laden appraisal" of policy. Though they cited specific social science information, "the final decision whether or not to proceed with a particular policy, was more likely to depend upon an appraisal of 'soft' knowledge (nonresearch based, qualitative and couched in lay language). . . ." These officials were also eclectic in their use of information sources, relying on newspapers, TV and popular magazines as well as on scientific government research reports and scientific journals. Caplan got "the overall impression that social science knowledge, 'hard' or 'soft', is treated as news by these respondents — allowing its users to feel that their awareness of contemporary social reality does not lag behind." (p. 231)

3. Because a policy maker is often confronted with "an overwhelming number of bewildering and complex responsibilities," research is often sponsored to help him "find his way out of this conceptual mudhole." Unfortunately, the purpose of such research is, according to Caplan, "rarely made explicit to the researcher." Some interviewees, for example, supported the use of social indicators and they even named some. But when asked about the uses they would make of such data, "The responses were so rambling and diverse that it was impossible to derive empirically based coding categories for purposes of quantification." Caplan stresses here a precondition for the conduct of evaluation and similar studies which is seemingly crucial: there must be some "previously agreed notion of what purposes" are to be served by the expected results of the study. (pp. 231-232)

4. The final general conclusion by Caplan is that utilization of study results is more likely if:

> a. they (findings) are not counter-intuitive;
> b. they are believable on the grounds of objectivity; and
> c. their action implications are politically feasible. (p. 232)

Caplan points out that "objectivity" may relate both to methodology and to interpretation. He suggests that "perhaps more than for other reasons, careless, irresponsible and shoddy program evaluations were cited by respondents to discredit social science research." He notes that "The ultimate test of data acceptability is political. Rarely are data in their own right of such compelling force as to override their political significance. This is an ancient issue and much has been written on it; it remains important."

In concluding his analysis, Caplan observes that the conditions which appeared in his study to influence utilization overlap and appear to be "somewhat contradictory."

> It does appear, however, that the major problems that hamper utilization are *nontechnical*. That is, the level of knowledge utilization is not so much the result of the slow flow of relevant and valid knowledge from knowledge producers to policy makers, but is due more to factors involving values, ideology and decision-making styles. (p. 233)

Comments:

This study identifies the orientation of an official as an influence on the types of knowledge that are sought and used. It does not, however, define *use*, *utilization*, or *impact*. It apparently accepts the self-reports of respondents. Is use merely reading a report? Or is it a change in the understanding of the reader? Or is it an action which would not have been taken in the absence of a study? Or something else? The possible alternative meanings of "use" leave us guessing about some of the implications of the Caplan survey.

Second, it is not clear from this reporting what proportion of those with an "academic orientation" (the most frequent users of "social science knowledge"), for example, were in research, analysis and evaluation positions in which the nature of their jobs and roles required the use of social science sources.

Third, the study underscores the fact that the overwhelming majority of officials use multiple sources and types of information and that the dominant use of information is political.

Finally, Caplan could not conclude *from this study* that the "relevance and validity" of knowledge does not inhibit its use. He seems to believe that there is an adequate flow of valid and "objective" social science knowledge relevant to many (most?) policy problems.

Assessment of State and Local Government Evaluation Practices in Human Services. (Baumheier, Edward C., et al. Denver: Center for Social Research and Development, February 1977.)

The Center for Social Research and Development, University of Denver, conducted this study of evaluation practices for the Office of the Assistant Secretary for Planning and Evaluation, DHEW. The purpose of the study was to assess the evaluation practices of State and local governments in areas of human services and to provide these governments with critical assessments of "various organizational structures, methodological techniques, and operational procedures for conducting and utilizing program evaluations."

Three-day visits were made to nine States "selected as good examples ["exemplary"] of evaluation units located in a wide variety of organizational structures within State and local governments." The sites included evaluation units in the Department of Health and Rehabilitative Services, Florida; Department of Public Welfare, Texas; San Diego County, Calif.; Human Resources

Administration, New York City; Hennepin County Mental Health Center, Minneapolis, Minn.; Vocational Rehabilitation Service, Lansing, Mich.; Office of the Governor, State of Washington; and the Joint Legislative and Review Commission, Commonwealth of Virginia.

The main conclusion of this study is that "decision makers at all levels benefit by having an evaluation unit available to them to provide specific information about programs in their areas of responsibility." Though too numerous to reiterate in full, here are several of the study's specific findings:

> Some evaluations were initiated to identify program problems, others to justify the value of a program, and a few without a clear purpose in mind. Sources of initiation included the legislature, the executive, the program to be evaluated and the evaluation unit itself.
>
> Activities identified as "evaluation" took many forms. The two most common were (a) *evaluation research* which tended to follow experimental methodology, and to be *outcome* oriented, summative in nature and limited in scope; and (b) *performance monitoring* which tended to be formative evaluation of the service delivery *process* and descriptive rather than experimental.
>
> Performance monitoring was found more prevalent, addressed more practical concerns, and was utilized to a greater extent than evaluation "research." Study recommendations suggest that experimental research be left to the Federal Government, while States and localities pursue performance monitoring.
>
> "The evaluation units that worked the most actively to promote utilization were the units whose evaluations were the most utilized." They sought approval from decision makers for their recommendations, developed plans for implementing recommendations, provided technical assistance for implementation, and checked periodically on progress toward implementation.
>
> Three conclusions were reached about the transferability of evaluation activities:
>
> > First, none of the specific findings of the evaluation case studies are directly transferable to other settings. This is true because no two human service programs are alike. Even categorical programs are administered in widely divergent fashions across the country. . . .
> >
> > Second, few of the specific evaluation methodologies utilized in the case studies are directly transferable to other settings. This is true because performance monitoring does not follow as structured a set of procedures as experimental research. . . .
> >
> > Third, the general experience of the sites in establishing and operating evaluation systems are clearly transferable to other settings. . . . (pp. 4-5)

Experience in individual sites is recounted in a set of nine case studies which accompany the main report.

Comments:

This is one of the few studies which attempts to describe what local and State evaluation units are actually doing in the name of evaluation and with what degree of perceived success. The site reports are worth the time of those who want to establish a new or strengthen an existing evaluation capability.

The criterion of evaluation impact used in the study was the combined *judgment* of the evaluation unit and the field researcher. Together they selected one evaluation study of apparent high impact and one of relatively low impact and then examined the factors seemingly associated with each. It is unclear how the resulting nine studies of high impact and the nine of low impact compare with the dozens of others carried out by the subject evaluation units.

The study tends to confirm the general conclusions reached elsewhere that the context of programs varies widely, that specific evaluation methods and findings cannot be transferred wholesale from place to place and problem to problem, and that evaluation approaches and methods

have to be tailored to specific programs and problems in specific contexts. The case studies indicate that the performance monitoring recommended by this study consists of *descriptive* studies of program *and* management *practices* and *processes*. The studies clearly are not "scientific" evaluation research on the "causes" of program effects. They fit more closely the traditional mold of organization, operations, and management studies and analysis, the kind that State and local program leadership apparently find the most desirable and useful.

Interim Analysis of 200 Evaluations of Criminal Justice. (Larson, Richard C., et al. Cambridge: Operations Research Center, Massachusetts Institute of Technology, May 1979.)

This is one part of a larger study of methods used in criminal justice evaluations. It is based on a "structured sample" of 200 of the "best" evaluations selected from among roughly 1,500 studies identified as evaluations in the National Criminal Justice Reference Service (NCJRS) in late 1977. Fifty percent of the sample was intentionally selected from "logistical" programs in which "the movement of persons, material or other entities was an important element." The other 50 percent came primarily from "social service type programs in which counseling or some other type of service is provided to one or more client groups." The sample also reflected three different Law Enforcement Assistance Administration (LEAA) evaluation efforts: evaluations of information in an area (police preventive patrol); exemplary projects nominated for wider replication; and LEAA "anti-crime impact cities" programs. The sample also focused on studies which "purported" to use "certain current methodologies" such as time series analysis, experimental design, models, decision analysis, etc. (pp. 5-7)

About 1,500 NCJRS document summaries were reviewed and graded subjectively on a scale from A to D. Those studies with the "highest grades" (the "best") were selected. Readers spent roughly 4 hours with each evaluation report and completed a checklist of 31 entries to "obtain information regarding evaluation input, process, and outcome, and to assess in a general way the relevance of the methodology employed, and the quality of the documentation." (pp. 10-15)

The study team notes that they were concerned with the *use* of evaluations by decision makers, the likely *value* of the evaluation information generated, the misuse and abuse of quantitative methods and "the use of adaptive evaluation methods to respond to feedback from the field." Because adequate information on *use* was not available in the documentation examined, however, the team is administering additional questionnaires to evaluators and "consumers" of evaluation reports. They hope to document the "budgeting, timing, planning and design of evaluation (inputs), interaction between program staff and evaluators, e.g., communication (process), and the ultimate use of the evaluation." This interim report contains many summary descriptive statements about the information provided in the evaluation documents examined. Many relate to issues of technical methodology which are not our concern. Only a few of the two and one-half pages of tentative conclusions (pp. 68-70) are of interest here.

> Target population was not discussed in one-third of the sample. "A slight majority of reports did not consider whether the program had been implemented as designed, and description of program activities is frequently inadequate as well."
>
> Experimental and quasi-experimental designs were the most common types, followed by narrative case studies; there was little use of statistical or formal *models*.
>
> "The most widespread problems were misapplication of common statistical techniques and difficulties in attributing outcomes to program activities; i.e., poor choice of performance measures.
>
> "There is a generalized lack of documentation of data collection procedures, and data were sometimes poorly used once obtained. A complementary problem is poor presentation, more so in qualitative than in quantitative studies." (pp. 68-69)

The report concludes that many of the problems identified "are manifestations of the basic problem with the criminal justice evaluations in our sample, namely that quite frequently the evaluation methodology used is not well matched to the type of program being evaluated." (p. 69)

The report recommends the use of:

> ...well-structured hypotheses or mental models concerning how the program should work. It is very important that the evaluator have some notion of how progam activities are linked to desired outputs and to other social, economic and political activities in the subject community: in many instances, the use of statistical or other formal models would help immensely. The point of stressing the need for articulated hypotheses is to wean evaluators away from the textbook formulas to which they were taught to adhere with little regard for circumstances. (p. 69)

In addition, "difficulties in applying various types of social science methods and measures were frequently manifested.... Common sense occasionally gets lost in the pursuit of elegant methods." (p. 70) The study team is pursuing better documentation of the input, process and outcome characteristics of the 200 evaluations.

Comments:

This report is primarily oriented to formal technical (social science research) issues related to study design, formal methods, data use, etc. Evidence already presented in this monograph suggests that the emphasis on formal methods has been grossly exaggerated. The usefulness and impact of evaluations seem related more closely to the articulated, situational and felt information needs and options of intended users and decision makers. The *qualitative* commentary in this report seems to bear out the point. This interim study seems to rest partly on the dubious assumption that the more formal the methods employed, the more useful the evaluation. Beyond some minimal level of credible methods, this assumption is doubtful. The next two studies provide additional reasons.

"Lessons Learned From Federally Mandated Program Evaluation for Community Mental Health Centers: Framework for a New Policy." (Flaherty, Eugenie Walsh, and Windle, Charles D. Submitted to *Evaluation and Program Planning*, May 1980.)

This paper examines assumptions that appear to underlie the extensive program evaluation requirements of the Community Mental Health Centers (CMHC) Amendments of 1975 (P.L. 94-63). It discusses four alternative evaluation models and their "sometimes contradictory purposes" and the conflicting motivations and values about evaluation held by key parties in evaluation. The authors then propose nine "principles" to guide future Federal CMHC evaluation policy and suggest ways to guide policy on accountability and program improvement.

This study is one of the few apparent attempts to examine critically the experience with a set of federally mandated program evaluation requirements for a *specific program* and to infer lessons and guidance from that experience. It draws on a wide variety of evidence and experience including a 1978 study by Flaherty and Olsen of evaluation in nine CMHC's funded by the National Institute of Mental Health (NIMH) and conducted by the Philadelphia Health Management Corporation.

Several of the authors' observations and conclusions are of interest. They cite, for example, the findings of three studies and conclude that, Federal fears aside, CMHC's would continue to do some evaluation work even if Federal requirements were removed. Centers would reportedly reduce the amount of evaluation by eliminating activities that are not "stimulated by center need." They doubt that program self-evaluation will contain costs and report that "program evaluation generally has been used to justify program expansion rather than program contraction." They also conclude that the "stringent evaluation requirements in P.L. 94-63 were based on assumptions that are only unevenly supported by available evidence and analysis" and "may not be justified." (p.3)

The authors identify four alternative models of evaluation (amelioration, accountability, advocacy and traditional research) and conclude that there is "little evidence" on which purposes and use

evaluation is actually put to in CMHC's. They conclude that "external pressure to do evaluation is associated with minimal utilization; only when evaluation is initiated because of center's own felt need is evaluation judged very useful (Flaherty and Olsen, 1978)." (p. 7)

The authors judge the use of evaluation for advocacy purposes "of doubtful integrity and long-run value for improving the quality of mental health services, although it has some immediate value for program viability." (p. 8) They also believe that the "traditional research model" of evaluation conflicts with the other three models partly because:

> . . . it is likely to displace these applied forms of research, because it is of more interest and personal value to program evaluators, thereby shifting the topics, approaches and funds away from program relevance and use. (p. 10)

Centers apparently comply only "minimally" with a requirement for an annual evaluation report for citizens. Few mechanisms for communication between centers and citizens exist and lack of compliance apparently springs "most importantly" from a "lack of citizen pressure, knowledge, or interest (Flaherty and Olsen, 1978)." (p. 8) The authors conclude that "These four models of evaluation are incompatible." (p. 9)

Flaherty and Windle take the reader through a parallel discussion of the varied and often conflicting "motivations and value systems" of "key parties in evaluation": center administrators, clinicians, citizens, service consumers and evaluators. Evaluation is most beneficial to administrators when it can be used, alternatively, to satisfy external requirements, describe the center to outside groups, assist management decision making, bring prestige and respect as evidence of serious efforts at self-management, increase the administrators' control of staff or visibly increase their ability to improve the center. (pp. 20-21) The authors summarize:

> These benefits occur most immediately when evaluation is conducted under the Advocacy Model, and next most quickly under the Accountability Model directed at funding and governing agencies but not at citizens. Benefit is most delayed and diluted in impact when evaluation is conducted under the Amelioration and Traditional Research Models, which take long to generate findings and are uncertain in results. (p.21)

Finally, the authors note that several studies suggest that the Community Mental Health Center Amendments of 1975 require evaluation "far in excess of centers' capacity and resources." (p. 22)

Flaherty and Windle derive nine "principles" for Federal policy for CMHC evaluation. Paraphrased and in summary form, they appear to suggest that evaluation requirements should:

- Be *feasible* and "not exceed by much the capacities of agencies to comply."
- Be *flexible* to accommodate varying programs' processes, and evaluation topics, purposes and methods, and to permit discretion about what, when and how to evaluate.
- *Focus* on accountability to the public and be limited to a few issues of importance, especially *descriptions* of what was accomplished and not program judgments about what was done.
- *Not* require "studies of client outcome" that are too expensive and complex and should be left instead to "special research."
- View evaluation as *developmental* and not require uniform and standard evaluation activities from programs at many different stages of development.

In addition, requirements should safeguard the confidentiality and dignity of program clients and staff, provide for routine dissemination and publicity of evaluation results, provide for evaluation of the evaluation activities themselves, and provide independent support for citizen participation in evaluation. (pp. 23-27)

Comments:

This paper is worthwhile reading for the lessons it conveys to anyone considering mandating evaluation requirements from higher to lower levels of program and/or government. It adds additional

weight to the view that outcome studies should not be routinely mandated of local programs and reinforces the position that evaluation serves local programs best when it satisfies *locally defined* purposes and uses.

Utilization-Focused Evaluation. (Patton, Michael Quinn. Beverly Hills, Calif.: Sage Publications, 1978.)

This is a wide-ranging, well-illustrated and probing discussion of evaluation. Of interest here is a study of the utilization of 20 Federal health evaluations that serves as part of the basis for Patton's proposed practices to increase the likelihood that evaluation results will be utilized.

In the fall of 1975, Patton and participants in an evaluation methodology training program at the University of Minnesota conducted inductive followup case studies of 20 "examples of excellence" in national health evaluations "selected from among 170 evaluations on file in the Office of Health Evaluation, DHEW." (Less than half the 170 studies qualified as "evaluation research" since many were found to be "nonempirical think pieces or policy research studies aimed at social indicators in general rather than evaluation of specific programs.") The 20 evaluations included 4 mental health center activities, 4 health training programs, 2 national assessments of laboratory proficiency, 2 of neighborhood health center programs, 2 studies of health services delivery systems programs, 1 alcoholism training program, 1 health regulatory program, 1 Federal loan forgiveness program, 1 training workshop evaluation, and 2 evaluations of specialized health facilities. Six of the 20 cases were internal evaluations, 13 were conducted by outside groups and 1 was done by one Federal unit for another. They ranged from a one-person 3-week program review to a 4-year evaluation which cost 1.5 million dollars.

Three "key informants" were intensively interviewed about the utilization of each of the 20 cases: the study project officer, the person identified by the project officer as the decision maker for the program or the person most knowledgeable about the study's impact, and the responsible evaluator. Most of the decision makers were office directors (and deputies), division heads or bureau chiefs. Interviews averaged 2 hours and ranged from 1 to 6. They were taped and transcribed. Three staff members independently analyzed the transcriptions for patterns and themes. Hypotheses were formulated and interviews were re-examined for relevant evidence, pro and con.

Interviewees were permitted to define impact *in their own terms* for these exemplary evaluations. Seventy-eight percent of the decision makers and ninety percent of the evaluators felt that the evaluation had had an impact on the program. Eighty and seventy percent, respectively, felt there were also "non-program" impacts. Perceived impacts were not, however, the kind where new evaluation findings "led directly and immediately to the making of major, concrete program decisions." Patton reports:

> The kind of impact we found, then, was that evaluation research provided some additional information that was judged and used in the context of other available information to help reduce the unknowns in the making of difficult decisions. The impact ranged from "it sort of confirmed our impressions. . . confirming some other anecdotal information or impression that we had" (DM 209: 7,1) to providing a new awareness carrying over into other programs. . . . (p. 30)
>
> . . . utilization is a diffuse and gradual process of reducing decision-maker uncertainty within an existing social context (cf. Levine and Levine, 1977). (p. 34)

Patton concludes that utilization of evaluation studies can be increased and better targeted but that the results will be more modest than rationalizing decision-making processes.

Throughout the book, Patton painstakingly reiterates that the touchstone of an evaluation that is likely to be useful is *not* the *evaluator's* theories, methods, specification or interpretation of program goals or evaluation criteria, but rather:

> *The first step in the utilization-focused approach to evaluation is* IDENTIFICATION AND ORGANIZATION OF RELEVANT DECISIONMAKERS FOR AND INFORMATION USERS OF THE EVALUATION. (Emphasis in the original, p. 61.)

Patton stresses the importance of what he calls "the personal factor," which emerged unexpectedly in the study of 20 health evaluations.

> To target an evaluation at the information needs of a specific person or at a group of identifiable and interacting persons is quite different from what is usually referred to as "identifying the audience" for an evaluation. Audiences are amorphous, anonymous entities. Nor is it sufficient to identify an agency or organization as recipient of the evaluation report. Organizations are an impersonal collection of hierarchical positions. People, not organizations, use evaluation information. (p. 63)

He reiterates:

> The specifics vary from case to case but the pattern is markedly clear: where the personal factor emerges, where some individual takes direct, personal responsibility for getting information to the right people, evaluations have an impact. Where the personal factor is absent, there is a marked absence of impact. Utilization is not simply determined by some configuration of abstract factors; it is determined in large part by real, live, caring human beings. (p. 69)

In the last chapter, Patton summarizes "utilization-focused evaluation:"

> There are only two fundamental requirements in this approach: everything else is a matter for negotiation, adaptation, selection, and matching. First, relevant decisionmakers and information users must be identified and organized — real, visible, specific and caring human beings, not ephemeral, general and abstract "audiences," organizations, or agencies. Second, evaluators must work actively, reactively and adaptively with these identified decisionmakers and information users to make all other decisions about the evaluation—decisions about research focus, design, methods, analysis, interpretation, and dissemination. (p. 284)

Between the summary of his study of 20 evaluations and the closing chapter, Patton takes the reader through a wide array of issues, illustrations, study evidence, theory, anecdotes, personal experiences and basic topics including "focusing the evaluation question," "the goals clarification game," "the methodology dragon," etc.

Comments:

This is a pragmatic, realistic, carefully stated and broadly based discussion of evaluation. It is laced with the lessons of experience and common sense and is highly recommended. Much of Patton's advice is similar to or consistent with the guidance given in part VI of this monograph.

Other Studies

Our search identified a few other studies, usually funded by the Federal Government, that examined State-level evaluation activities either as a single focus or as part of a broader look at program management. Typically, however, these studies appear to accept evaluation activities at face value. They do not explore use or impact, but nonetheless conclude by urging *more* evaluation.

For example, Pacific Consultants (February 1977) made site visits to eight States and one Federal Region and surveyed the remaining States by phone. In this study of social service evaluation under Title XX sponsored by the Social and Rehabilitation Service of DHEW, "level of evaluation performance" was indicated by the "number of studies completed or in progress." High performers were defined as States with 6 to 18 evaluation studies; moderate performers with 1 to 3 studies; and low performers with no studies. An examination of the 6 high performer States suggested that they *tended* to focus on impact studies; identified "program planning and improvement" as the primary purpose of evaluation; had planned substantially for social services; had special evaluation units that were "broad-scope" and relatively large (nine or more full-time equivalent staff); and had at least $150,000 available for evaluation. The study cautions that the descriptive factors they examined were not fully explanatory, and that "a number of factors included in the model must *coalesce within the same state* to produce significant probability of high performance." (Emphasis in the original, p. 16.) The study also identified 81 evaluation studies that were either completed or in progress: 17 management, 11 client characteristics, 34 process and 19 impact. The contractor saw an increase (at least short-run) in the number of process and impact studies.

While this study concludes that there is an overall improvement in the state of the art of evaluation since the implementation of Title XX, no attempt was made to assess the utility, impact or use of the

studies already completed or the conditions associated with that use. Here, as commonly elsewhere in the evaluation literature, the general value of formal evaluation studies is taken for granted and the dubious assumption often made that the more social scientific the better (usually in terms of traditional methodological characteristics). Oddly, these studies, which are often cloaked in the semblance of "science," appear to rest on circular reasoning and do not explore or sometimes even raise the basic questions: Of what value, worth, use, impact or relevance were these studies? To whom? Compared with what?

Finally, the Urban Institute conducted a 2-year study of State implementation of Federal Title XX social service programs. (Benton, Feild and Millar, 1978) This study also reported that there was "optimism" among State-level interviewees that "the use of evaluation data would increase over the next 3 years." More self-consciously than some other studies, however, this one at least questioned the assumption that producing more evaluation data will result in its use in decision making processes.

There are surely other studies of the use and impact of program evaluation that have been overlooked. Readers acquainted with them are urged to add the evidence to what has been presented here and come to their own conclusions.

General Conclusion

In an attempt to uncover evidence on the actual use, utility and impact of formal program evaluation, we searched for and screened a large volume of documents and studies. We did not find a body of valid, scientifically verified evidence which upholds the many claims for the value of formal evaluation. We found, instead, about a dozen or so assorted studies that bear on this issue and selectively summarized them. On the whole, they suggest a small, uneven, and modest use and impact of formal evaluation studies as these studies have been initiated, designed and carried out in the past. They also point to some practical tips for the practitioner.

Drawing on this eclectic body of evidence, the testimony of experienced evaluators, discussions with experts and our own experience, we give in part VI some general guidance, suggestions and rules-of-thumb that might help the State and local agency official, manager and practitioner decide what to do when confronted with decisions about conducting formal evaluation. Before identifying some of the proposed reforms to traditional evaluation theory and practice, we suggest a few general lessons that the Federal experience with program evaluation might suggest to other levels of government.

Intergovernmental Lessons

The U.S. system of federalism provides opportunities for trial and error and for cumulating experience with an approach in a circumscribed way short of universal adoption or application. In principle at least, learning from these experiences may be transmitted to other levels and locations in the system. One part of the social or governmental system may then learn from the successes or failures of another. These learning experiences are possible and have occurred historically in multilateral directions (many from State and local levels to the Federal level). Some intergovernmental and intersector borrowing of practices, however, do not appear to be based on learning but rather on copying and mimicry. In these instances, untested claims for an approach may continue to run well ahead of the caveats of experience. The Federal Government may have cycled through the adoption, use and adaptation of an approach like PPB. The States, by contrast, may be starting a cycle with the same premises that the Federal Government may have already abandoned or modified.

History may be at or beyond a similar point in the case of program evaluation. Federal expectations and practice are much more modest now than in the late sixties or early seventies; claims have been substantially muted by the force of experience. Yet reports suggest that some States have begun to copy not the recent but the earlier Federal experience without the benefit of the lessons already learned by the Federal Government. What, then, are some of the lessons about the use and practice of program evaluation that might contribute to a satisfactory intergovernmental learning experience? Here are some that appear to transcend the operational suggestions given later in part VI.

1. Be selective in the requirements for and use of formal program evaluation. Do *not* mandate program and project evaluation requirements (through laws, regulations and other rulemaking) uniformly and comprehensively for every program. This will lead inevitably to redundancy, waste and the diversion of some resources from more useful management purposes. Not every program can or should be evaluated formally.

2. Do not expect that formal program evaluation will yield satisfactory overall conclusions about "all-or-none" questions or about the overall worth and value of programs. These judgments emerge from social and political processes and not from formal studies.

3. Do not mandate *outcome* evaluation studies. They are expensive, complex and often impossible. These efforts are best left to special applied research, probably conducted most reasonably on the national level.

4. Do not mandate any single evaluation methodology, ideology, approach or method. Appropriate tools and approaches for formal evaluation should be suited and fitted to specific program problems and to the information needs of a wide variety of agency administrators, program officials and other influentials whose circumstances differ substantially.

5. View evaluation in a broad sense as *study* (rather than formal research) and include management, policy, operations, procedural and workforce efficiency and effectiveness studies.

6. If program evaluation requirements are to be established, make them *selective, restrained, permissive* and *enabling*. They should not be universal, ambitious, restricting, detailed and compulsory.

7. Do not mandate evaluation because it will be good for the other guy. If the officials of a State or local agency do not intend to use the results of evaluation in concrete ways, they should not mandate it for others.

8. Be modest in expectations about the payoff of formal evaluation for resource allocation and program management decision making.

9. Keep program evaluation in perspective. Reexamine the evidence about its likely payoff. Think about it.

V. Proposed Reforms of Traditional Formal Evaluation

Practical experience with formalized program evaluation and its seemingly small payoff has led a variety of self-conscious observers, practitioners and commentators to propose reforms for the theory, practice and role of evaluation. Though some of these reforms may have no immediate practical implication for the State and local practitioner, they reflect how profoundly the area of formal program evaluation is under reconsideration and transformation. Too numerous to treat in number or detail, here are a few of the dominant proposed reforms.

Sustain a Reasonable Measure of Self-Evaluation

Past evaluation philosophy has emphasized that the "best" evaluation is done from outside the program, either at a higher level in an organization, or by a body, group or institution beyond the direct influence of the program to be evaluated. This advice appears to be based on the joint premises that (a) agencies and programs left to self-evaluation will be self-serving and biased ("they can't be trusted"), and (b) those outside will have no vested interest and will be more impartial and unbiased. Both these premises may be faulty. First, there is no pure unbiased, or value-free evaluation. There are only many perspectives from which different value judgments may be made, some more persuasively than others. Outside judgments are not always more compelling than those inside. Second, holding program officials responsible and accountable for a program and its performance should entail giving them some share of the responsibility, encouragement and resources for self-evaluation.

In general, using mixed approaches of both inside and outside evaluation seem more sensible than using either one exclusively. It is significant that while criticism and pressure for program reform may come from outside, many reforms can only be effected by those inside. Reforms that find their source partly on the inside may occur more acceptably, more effectively and more enduringly than those "invented elsewhere." This position urges restoring the respectability of internal or self-evaluation in combination with other varieties.

Support Competitive Evaluations

This reform proposal can be viewed as an extension of the first one. It acknowledges that all individual evaluations will be partial, spring from some value position and be without much external cross-checking. To increase the range of both analytical and value input, several competitive evaluations of the same program are urged. Out of this competitive adversarial process will come, it is argued, better cross-checking and error-correction than is possible with a single try. This position appears to rest on the general logic that underlies adversarial judicial proceedings, competitive markets, and much of science. (Polanyi, 1964; Toulmin, 1972; Fleck, 1979; Judson, 1979) A practical implication of this proposal is risk-spreading mentioned earlier: do several small evaluations of different program dimensions or problems rather than one intended to be global or comprehensive. The competitive evaluation position also urges that multiple evaluations of program activities consciously reflect major alternative views of or positions on program problems and issues. In over-simplified terms, one evaluation might be undertaken by a provider-oriented group, another by a

client-oriented group and a third by a finance/management-oriented group. Or, one evaluation might examine component A of a given program, another component B, and still another component C. In still another situation, a given program or one of its components might be evaluated simultaneously from two or three competitive political or ideological positions. It is presumably as a result of competitive evaluation and policy analysis that more reliable and relevant information and remedies would emerge.

Improve Citizen and Client Participation in Program Evaluation

This proposal, a variant of competitive evaluation, is based on the fact that most resources and responsibility for existing formal public program evaluation now lie within the control of executive agencies. Formal evaluation efforts of these agencies are, it is reasonably suggested, heavily influenced by motives of self-maintenance and stability. They also tend frequently to be oriented toward existing service provider arrangements, affiliated organizations, and professional groups and associations, and toward dominant existing commercial and financial interests. Amidst the din from these politically active program stakeholders, the voices of the client and the citizen-taxpayer are often muffled, if not lost. Though appropriate detailed mechanisms are not clear, the intent of this proposal is to increase the role of citizens in program evaluation. One specific recommendation is to make program evaluation results more accessible to the public. A stronger proposal is to make some share of evaluation funds and resources directly available to citizen and client-oriented organizations and associations. (Flaherty and Windle, May 1980)

Re-Examine Traditional Evaluation Premises

This reform proposal calls for a re-examination of the "fit" between (a) formal "rational/scientific" modes of information gathering and knowledge building and (b) the problem solving and program evaluation tasks that actually confront real-world operating agencies and programs. McLaughlin (1980) believes that "efforts to 'fix' existing evaluation paradigms [the experimental and input-output models] are unlikely to be fruitful." She concludes that (a) "many of the important factors in the local process of change may be inherently unquantifiable and not amenable to control," (b) "the logic of inquiry is wrong," and (c) "fundamental incongruence between the set of relationships presumed by our current logic of inquiry and the local reality has led to spending much time and energy in developing new instruments to measure outcome and calibrate inputs. These efforts typically are undertaken at the expense of rethinking the conceptual framework for learning from project experience." (pp. 45-46)

Lindblom and Cohen (1979) have also taken a fundamental and critical look not just at social science-based evaluation, but at the larger class of what they call "professional social inquiry." Similarly in England, Sharpe (1976) has critically examined the relationship between the social scientist and policy making.

Conduct a Pre-Evaluation or Feasibility Assessment

This reform has been developed most extensively by Wholey (1979) and Schmidt, Scanlon and Bell (1979). Wholey, for example, cautions an agency not to rush to intensive evaluation until it has gone through some preliminary or pre-evaluation steps. He suggests a "sequential purchase of information." In order of apparent increasing commitment of resources, the steps are: evaluability

assessment, rapid-feedback evaluation, performance monitoring and intensive evaluation. Wholey characterizes his proposed incremental sequence this way:

> Rather than proceed directly from the program to be evaluated to intensive evaluation of program effectiveness, we insert one, two or three preliminary evaluation steps, any one of which may produce sufficient information for policy or management decisions. . . . Our approach produces relatively inexpensive information on program performance — within months, rather than years. (pp. 13-14)

The next and final part provides guidance to the practitioner confronted with a decision about doing formal evaluation.

VI. Evaluating the Expected Value of Doing a Formal Evaluation

Why Carry Out a Formal Evaluation Activity?

There are several possible alternative purposes: compliance evaluation, formal social research, miscellaneous purposes and problem-oriented evaluation.

Compliance Evaluation

In the last 15 years, governments have increased legal requirements in laws and regulations for formal organizational functions such as planning, needs assessment and evaluation. (Zangwill, 1977; Knezo, 1974) These mandated activities are often preconditions for new or continuing financial support. Many plans, needs assessments and evaluations, however, are created primarily for compliance purposes, and not primarily for the value they may have to a sponsoring agency (for example, National Institute of Mental Health, 1977; Kimmel, 1977; and Lovell et al., 1979). "Compliance evaluation" is likely to entail the minimum and sometimes symbolic effort required to achieve compliance. Government evaluation requirements may, however, allow a number of alternative evaluative and analytical activities. It may or may not be possible to generate benefit to the agency while still complying.

Formal Social Research

Some social scientists and professional evaluators apparently view the availability of funds for program evaluation as an opportunity for social research, somewhat independent of its payoff for policy and management purposes. This justification has been offered for work on social indicators and social surveys and for methodology development. One result of publicly funded evaluation studies in the past may have been a test of the utility and relevance of formal research methods applied directly to operational program issues and policy problems. The result seems to be that the fit between these two sets of activities is poor, perhaps even counter-productive. (Lindblom and Cohen, 1979) Some general social benefit, however, may have been derived (in the form of social learning) from exposing large numbers of researchers to the actual processes and complexities of social problem solving and policy formulation, and from simultaneously giving policymakers and program officials a better appreciation of both the possibilities and the limitations of formal research methods applied in a public policy setting.

Miscellaneous Purposes

Beyond compliance and research lie a broad range of other possible reasons (motives) for considering some form of evaluation activity:

1. To confirm what is already known or suspected about a program, either its weaknesses or strengths;
2. To stimulate political response to a program by pressuring it, generating legitimacy for it or stimulating further advocacy support for it;
3. To generate field feedback in the form of site-visit reports, case studies, program illustrations or descriptive information for use in agency program documents or justifications;

4. To contribute to a general background of information or "enlightenment" (e.g., Weiss, Fall 1977);
5. To emulate cosmetically the practices of "scientific management;"
6. To play out what Morrill and Francis (January 1979) identify as this "...syndrome: we have a problem, we don't know exactly what it is and don't have time to think it through, so let's get a study to figure it out " (p. 28); and
7. Other reasons and motives recognized by the reader.

The purist rational evaluator might object that some of these possible reasons (motives) for program evaluation are political and that evaluation should be "free of politics." The realist might respond that public agency evaluation that is free of all politics is likely to be free of all relevance.

Problem-Oriented Evaluation

This type of evaluation derives from felt problems and issues, the exploration, clarification and amelioration of which *may* be enhanced by some form of evaluative activity.

Remedies for these problems, often problems of management, process, procedure and practice, do not lie in the establishment of "scientific facts" through comprehensive research, but in a more circumscribed, pragmatic and problem-oriented mode of identifying issues, articulating their structure, and finding or inventing feasible and practical remedies to reduce or resolve them. For the exploration and remedy of these varied problems, no *single*, *universal* approach, method or tool exists beyond perhaps observation, thought, reflection, and common sense tutored by experience and trial and error. In State and local government settings, some issues and problems direct attention to the use of trouble-shooters, management analysis, operations analysis, descriptive studies, trend analyses (of costs, service utilization, staffing patterns, and so on), rapid feedback explorations and performance monitoring. Other problems point to a broad array of interactive problem-solving mechanisms. Some may benefit from both interactive and formal study approaches.

The next three sections present suggestions for initiating program evaluation activities. Summary guidance is first outlined in table 4.

General Guidance for Program Evaluation

Expectations

If your expectations about the payoff of formal program evaluation are very high, lower them. If you expect to derive "scientifically verified" facts and conclusions, you will be disappointed. You are more likely to be satisfied if you expect small and not large additions to your understanding of a given program, its problems and possible remedies; partial reality testing and not global confirmation (or refutation) of your beliefs and opinions; and a *supplement* to (sometimes small) rather than a *substitute* for the information, knowledge and feedback which already exists.

Bias

While you may be able to control willful bias and blatant value loading, all program evaluation and performance monitoring activity is selective, value influenced and value embedded. If there were no values guiding an evaluation, of what possible interest and use would it be?

Table 4
Guidance for Program Evaluation

A. *Motives for Evaluation:*

Acknowledge the many possible alternative motives for evaluation. Decide which ones suit the immediate situation.

B. *General Guidance:*

1. *Expectations:* Be realistic. Keep them moderate.
2. *Bias:* Control what you can and be alert to what you cannot.
3. *Scale:* Break potentially large studies into several small ones.
4. *Abstractness:* Aim studies at concrete well-defined issues.
5. *Beneficial Interactions:* Maintain moderate levels of regular interaction among sponsors/users and evaluators.
6. *Risk:* Reduce risk of study failure — by spreading it.
7. *Politics:* Expect them and make the most sensible use of them.

C. *Pre-Evaluation Preparations (Homework):*

1. Identify specific, concrete program problems and issues; consider a "program evaluation issue paper."
2. Identify and interact regularly with expected users.
3. "Scout around" to get some feel for evaluation possiblities.
4. Consider several *alternative* types of possible evaluation:
   - "Quick and Dirty,"
   - Rapid Feedback,
   - Exploratory, or
   - Problem-Oriented.

D. *Useful Practices and Rules of Thumb:*

1. Fit tools to problems (and not vice-versa).
2. Know your evaluator(s).
3. Consider evaluation an interactive and negotiated process.
4. Do not isolate the evaluator(s).
5. Demand/prepare intelligible reports.
6. Ask evaluators to include qualitative reporting and judgments.
7. Keep evaluators involved in technical assistance.

Scale

A mix of several small program evaluation studies and activities directed at the same program are probably better than one large one. Studies of narrow scope are more likely to be focused, feasible and manageable. They are also more likely to pay off in terms of relevance, currency and cost (in both time and money).

Abstractness

Evaluation aimed at concrete, well-defined issues and areas is likely to be more useful, though maybe less dramatic, than open-ended studies guided only by an abstract interest in how well a program is serving "the public interest," meeting "comprehensive community needs," or achieving broad and diffuse goals and objectives.

Beneficial Interactions

A moderate degree of sustained interaction between evaluators and their sponsors (or intended users) is likely to result in better mutual understanding of the logic, possibilities and limitations of an evaluation; permit better tailoring of study scope, focus and method to felt problems, concerns and intended uses of evaluation results by agency personnel; induce a larger exchange of *qualitative* information; and reduce the surprise and potential threat of study findings.

Risk

Breaking a potentially large and comprehensive evaluation into smaller components is one way to reduce the risk of failure by spreading it. It is also a way to more easily fit tools, approaches and skills to varying dimensions of an evaluation problem. Similarly, it avoids putting all evaluation eggs in one methodological basket. This strategy was consciously employed at the Federal level by the National Institute of Education (NIE) when it answered a mandate from Congress for an evaluation of compensatory education programs. Study Director Hill reports that deadlines helped them spread risk:

> The deadlines also forced us to define simple projects that could be designed, put into the field, and reported quickly. We mounted a large number of small projects, each designed to accomplish a simple objective, rather than a few complex multi-purpose studies.... That practice had several advantages. It meant that each project was simple enough for one NIE staff member, rather than a team to monitor....
>
> Similarly, because our contractors did not need vast interdisciplinary teams of researchers, they experienced fewer managerial problems. Because projects were relatively self-contained, a problem or failure in one did not threaten the whole study.... We were, finally, able to conduct backup studies to protect ourselves against the possible failure of very crucial or difficult efforts....(Pincus, ed., 1980, p. 67)

Though this evaluation effort was large and lasted several years, the basic logic of risk-spreading also applies to small scale efforts.

Politics

Expect politics and make the most sensible use of them.

Additional advice can be found in many other sources (Patton, 1978; Flaherty and Windle, 1980; Levine and Williams, 1971; Morrill and Francis, 1979; Baumheier. et al., 1977).

Pre-Evaluation Preparations (Homework)

If an evaluation is under consideration, a simple sequence of thoughts and actions may enlighten the decision to proceed by exploring the purposes and uses an evaluation might serve. The sequence begins with two interrelated questions:

1. *What is (are) the specific issue(s) or problem(s) that an evaluation is intended to address?* This is probably one of the two most important questions which can be raised about a proposed evaluation. It is also one which may be ignored or skirted in the belief that evaluation is good in its own right. A detailed specification of the problem(s) or issue(s) that is (are) to be addressed is a prerequisite to judging whether an evaluation is appropriate. It represents essential homework. (Morrill and Francis, 1979)

One approach to elaborating program or policy problem(s) is to specify a set of questions to be answered. This can be done, obviously, "in the head" by a "thought experiment" — what would it be like if. . . ? (Wildavsky, 1979) or by some "back-of-the-envelope" jottings and calculations. A more structured approach is a program evaluation issue paper — a short written statement that lays out in tentative terms:

- The perceived nature and apparent structure of the program problem(s);
- Likely sources of the problem(s);
- Known and suspected evidence of the existence of the problem(s);
- Alternative actions that might be taken by the agency to remedy the problem(s);
- Indicators which might be employed to show progress toward resolving the problem(s);
- Estimated costs (of many kinds) and impacts of possible remedies for the perceived problem(s);
- Significant known or likely constraints on reducing the problem(s);
- Major evaluative, analytic or data problems that have to be faced and handled if further study and investigation is to proceed;
- A list of key steps in additional study or investigation that might be taken and an estimate of their cost, skill requirements and timing; and
- An identification of expected users of study results.

A program evaluation issue paper is a way to identify and describe the main features of a program problem(s) based on what is known or can be easily learned from existing sources. It is preliminary to more extensive evaluation, analysis or data collection. A well-developed issue paper should indicate whether a given program issue or problem can be clarified by further evaluation, analysis, better estimates of costs or impacts, a more refined understanding of the sources of a problem(s), or by some other action or response.

Results of this pre-evaluation should help indicate whether additional steps ought to be a management analysis, cost-effectiveness study, use of a trouble-shooter, an exploratory evaluation or some other action, mechanism or form of structured analytical work. Details of the contents and formats of two alternative issue papers, useful in both program analysis and program evaluation, are provided in Hatry et al. (1976). An intriguing case study of the "swine flu affair" and what a useful program/policy issue paper might look like the next time around have been prepared by Neustadt and Fineberg (1978).

A principal purpose served by clarifying and elaborating major program and policy issues and problems before starting more intensive evaluative work and data collection is to ensure that tools and approaches are selected to fit problems rather than vice-versa.

2. *Who are the expected likely users of evaluation results?* The answers to the first question and to this one are interdependent. As a growing number of experienced participant-observers have confirmed, relevant and useful evaluations do not grow out of idle curiosity, abstract concerns with science or the public interest or an academic interest in splitting intellectual hairs. They grow instead out of the live (sometimes nagging) questions, issues, problems and "felt" needs for information of involved and participating program leaders, managers, staff and other key program influentials.

As Patton (1978) has urged, *expected users of evaluation information should be identified early.* This should not be a guessing game. Interact with prospective users. Discuss a possible study with them. Elicit their questions and concerns *framed in their terms.* Indicate realistically what types of

information are likely to be generated, the probable quality of that information (including its limitations), timing, and so on. The skeleton of a program evaluation issue paper may be useful here. Do some additional legwork and then return for additional discussion with these potential users.

3. With a set of preliminary questions *and* likely users well in mind, have someone staff it out and scout around to get some preliminary feel for (a) the program's operations through interviews, site visits and examination of program reports; (b) the feasibility of carrying out the kind of inquiry you had in mind; (c) the level of effort, cost, timing and skills it might require; and (d) what might reasonably be expected to result. Scouting around may also contribute to the development of an issue paper.

4. If at this stage you decide to proceed with some variety of *formal* evaluation activity, consider the utility and value of a rapid feedback evaluation, an exploratory evaluation or "evaluability assessment." (Wholey, 1979; Schmidt et al., 1979) These further pre-evaluation steps may be more extensive than those discussed so far but they cover some of the same preliminary steps required by a full-blown formal evaluation. Details of one possible approach to an "evaluability assessment" can be found in the Schmidt et al. monograph published by Project SHARE, *Evaluability Assessment: Making Public Programs Work Better,* 1979.

As noted earlier, the formalization of evaluability assessment is relatively new. It was designed for use at the Federal level and is still in a developmental stage. It appears to grow partly out of the failure of traditional research modes of evaluation to pay off and partly out of the growing recognition over the past 5 to 10 years that under many circumstances a full formal evaluation will be neither feasible nor desirable. The principals who developed this approach report that there is no packaged experience available on its strengths, weaknesses or the conditions under which it pays off. They urge, as we do, restraint and caution in its use.

Useful Practices and Rules of Thumb

Fit Tools to Problems

Attempt to fit evaluation tools, methods and approaches to the nature and structure of perceived program problems rather than vice-versa. It is commonplace to find tools in search of problems and methods in search of applications.

Know Your Evaluator

Get to know your evaluator(s), their training, past work, style of thought, and preferred tools and approaches. Evaluators and other professionals are predisposed to do what they know best; i.e., their specialty. In a caricatured health-care analogy, surgeons cut, dentists drill, psychiatrists probe the mind and nutritionists explore eating habits. Where will the attention of your evaluator be drawn? To methods, to models, to tests and measures, to questionnaires, to interviews, to "gestalt" patterns, to qualitative considerations (like the history and context of the program), to philosophy? To input-output relationships, to the "black box" in between, to individual client outcomes, to program processes, to broad community impacts, to administrative and management mechanisms? To the political and bureaucratic environment? A reasonable exploration of the evaluator's predispositions, style of thinking and areas of professional comfort will permit more fruitful interaction between the client(s) and evaluator(s) and a more productive *negotiated* process of evaluation.

Consider Evaluation an Interactive Process

Do not deprive the evaluator of your concerns, the problems you perceive, areas of greater and lesser importance, blind spots, taboo issues, or unusual constraints on the agency and its range of possible corrective program actions. Sponsors of evaluation who do not articulate their concerns are not likely to get in return what they consider useful.

Similarly, encourage and ensure that evaluators interact with knowledgeable program officials and operatives not only at the start of a study but at regular intervals along the way. This will serve two purposes. First, it will permit the evaluator to access the qualitative and experiential knowledge, insights and understanding essential to *reality-based* evaluation. Much of this knowledge comes only from being involved in the historical development and daily business of program operations. No new evaluator(s) can approximate the collective wisdom and insights about a program of those who have been "dwelling" in it. Second, these interactions should (a) provide the evaluator with an enhanced understanding of the human roles and perspectives at work in the program; (b) increase the evaluator's knowledge of the details of the program's actual operations; (c) reduce the threat to and anxiety of the evaluator which may be associated with limited understanding of the program; and (d) forestall trips down technical alleys in search of answers that may be at the fingertips of the experienced program official.

Do Not Isolate the Evaluator

Do not isolate the evaluator with the misguided intention of protecting his or her objectivity or impartiality. Myths to the contrary, evaluation properly employed serves concrete purposes and interests and not abstract notions of science or the public interest. Encourage a flexible overall study approach that permits both the agency and the evaluator to suggest midcourse corrections. Within the boundaries of reasonableness and of prior commitments made to the evaluator, do not be reluctant to interfere in the course a study may take. Some agencies attempt to increase the articulation between their evolving interests and outside evaluators by considering the agency's evaluation monitor an integral part of the evaluation study team.

Demand/Prepare Intelligible Reports

It is a truism, regularly ignored, that findings of an evaluation that are not presented in an accessible and intelligible form will not be easily used. (Larson, 1979) In another place, the author and colleagues (Hatry et al., 1976) give some basic advice on the presentation of the results of program analysis. It applies with equal force to evaluation reports:

> Some of the most sophisticated and technically competent program analyses [Read "evaluations"] are unused and unusable. The reasons are varied: the main findings of the analysis may have vanished in a thicket of technical jargon; the recommended alternatives may be politically infeasible; the report on the analysis may have come too late; or the bureaucracy that must use the findings may be uninterested or resistant. In brief, program analysis [evaluation] can be elegant but irrelevant. (p. 9)

The authors further advise: have the report reviewed for technical quality and clarity; include minority reports; include a clear compact summary; acknowledge the limitations and assumptions of the study; use simple graphics to display major findings and conclusions; eliminate jargon; and tailor the presentation of results to the communication style of expected key users. (pp. 24-25)

Ask Evaluators for Qualitative Reporting and Judgments

Some of the most insightful and helpful reporting in evaluation studies may have little or nothing to do with the results of applying formal study methods. As implied earlier, evaluators usually come to the end of their formal methods before they come to the end of their wits. The observations and insights generated casually during the course of a study should be openly reported. One way to ensure this is to encourage evaluators to devote special sections of the report to qualitative reporting and personal interpretations.

Keep the Evaluator(s) Involved in Technical Assistance

For a variety of reasons, including the way some are employed (through time-limited contracts), evaluators may be "hit and run." But the presentation of study reports is but a beginning or midpoint of

program improvement. If remedial action is agreed upon by the agency, ensure that evaluators are, when possible, available to assist in effecting corrective action. This will not only sustain a blend of evaluator and program skills and knowledge, but also discourage the evaluator from formulating impractical recommendations that he/she may later have to help implement.

References and Selected Bibliography

Alkin, Marvin C.; Daillak, Richard; and White, Peter. *Using Evaluations: Does Evaluation Make a Difference?* Beverly Hills, Calif.: Sage Publications, 1979.

Almond, Gabriel A., and Genco, Stephen J. "Clouds, Clocks and the Study of Politics." *World Politics* 29 (1977).

Anderson, Charles W. "The Place of Principles in Policy Analysis." *The American Political Science Review* 73, no. 3 (1979): pp. 711-723.

Anderson, Scarvia B., and Ball, Samuel. *The Profession and Practice of Program Evaluation.* San Francisco: Jossey-Bass, Inc., 1978.

Banner, David K.; Doctors, Samuel I.; and Gordon, Andrew C. *The Politics of Social Program Evaluation.* Cambridge: Ballinger Publishing Company, 1975.

Baumheier, Edward C., et al. *Assessment of State and Local Government Evaluation Practices in Human Services.* Denver: Center for Social Research and Development, February 1977. (Contract Report HEW-100-76-0158 prepared for the Office of the Assistant Secretary for Planning and Evaluation, DHEW.)

Benton, Bill; Feild, Tracey; and Millar, Rhona. *Social Services – Federal Legislation vs. State Implementation.* Washington, D.C.: The Urban Institute, 1978.

Block, A. Harvey, and Richardson, David, Jr. *Developing a Client Based Feedback System for Improving Human Service Programs.* Human Services Monograph Series, no. 10. Rockville, Md.: Project SHARE, 1979.

Bronowski, J. *Magic, Science and Civilization.* New York: Columbia University Press, 1978.

Bronowski, Jacob. *The Origins of Knowledge and Imagination.* New Haven: Yale University Press, 1978.

Bruner, Jerome. *The Relevance of Education.* New York: W. W. Norton & Co., Inc., 1971.

Campbell, Donald T. "Assessing the Impact of Planned Social Change." *Evaluation and Program Planning* 2, no. 1 (1979).

Campbell, Donald T. "Reforms As Experiments." In *Readings in Evaluation Research,* edited by Francis G. Caro. New York: Russell Sage Foundation, 1977.

Campbell, D. T., and Stanley, J. C. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally, 1966.

Caplan, Nathan. "Factors Associated With Knowledge Use Among Federal Executives." *Policy Studies Journal* (Spring 1976).

Caplan, N.; Morrison, A.; and Stanbaugh, R. *The Use of Social Science Knowledge in Policy Decisions at the National Level.* Ann Arbor: Institute for Social Research, University of Michigan, 1975.

Caro, Francis G., ed. *Readings in Evaluation Research.* New York: Russell Sage Foundation, 1977.

Chelimsky, Eleanor. *An Analysis of the Proceedings of a Symposium on the Use of Evaluation by Federal Agencies: Volume 2.* McLean, Va.: Metrek, a division of the Mitre Corporation, July 1977.

Chelimsky, Eleanor, ed. *Proceedings of a Symposium on the Use of Evaluation by Federal Agencies: Volume 1.* McLean, Va.: Metrek, a division of the Mitre Corporation, March 1977.

Churchman, C. W., and Schainblatt, A. H. "PPB: How Can It Be Implemented?" *Public Administration Review* 29, no. 2 (1969).

Cohen, Alan B., and Cohodes, Donald R. *Assessment of an Approach to Evaluation Planning in Region V.* Cambridge: Urban Systems Research and Engineering, Inc., September 1977.

Conner, Ross F.; Rosener, Judy B.; and Weeks, Edward C. *Program Evaluation Within California State Agencies: An Assessment.* Irvine, Calif.: Public Policy Research Organization, University of California, May 1976.

Cook, T. D., and Campbell, D. T. "The Design and Conduct of Quasi-Experiments and True Experiments in Field Settings." In *Handbook of Industrial and Organization Research*, edited by M. D. Dunnette and J. P. Campbell. Chicago: Rand McNally, 1975.

Cornuelle, Richard. *Demanaging America: The Final Revolution.* New York: Vintage Books, 1976.

Datta, Lois-ellin, and Perloff, Robert. *Improving Evaluations.* Beverly Hills, Calif.: Sage Publications, 1979.

Demone, Harold W., Jr. *Stimulating Human Services Reform.* Human Services Monograph Series, no. 8. Rockville, Md.: Project SHARE, June 1978.

Dobos, Rene, and Escande, Jean-Paul. *Quest: Reflections on Medicine, Science and Humanity.* New York: Harcourt Brace Jovanovich, 1979.

Eaton, Joseph W. "Symbolic and Substantive Evaluation Research." *Administrative Science Quarterly* 6, no. 4 (1962).

Eckholm, Erik P. *The Picture of Health.* New York: W. W. Norton & Co., 1977.

Evaluation. Minneapolis: Program Evaluation Resource Center, Minneapolis Medical Research Foundation, Inc., 1972 to present.

Experiences in Evaluating Human Services. Human Services Bibliography Series. Rockville, Md.: Project SHARE, November 1977.

Flaherty, Eugenie Walsh, and Olsen, Kristin. *An Assessment of the Utility of Federally Required Program Evaluation in Community Mental Health Centers, Volumes I, II, and III.* Philadelphia: Philadelphia Health Management Corp., December 1978.

Flaherty, Eugenie Walsh, and Windle, Charles D. "Lessons Learned From Federally Mandated Program Evaluation for Community Mental Health Centers: Framework for a New Policy." Submitted to *Evaluation and Program Planning*, May 1980.

Fleck, Ludwig. *Genesis and Development of a Scientific Fact.* Chicago: The University of Chicago Press, 1979.

Freeman, Howard. "The Present Status of Evaluation Research." In *Evaluation Studies Review Annual, Volume 2*, edited by Marcia Guttentag and Shalom Saar. Beverly Hills, Calif.: Sage Publications, 1977.

Gamel, Nona N., et al. *State ESEA Title I Reports: Review and Analysis of Past Reports and Development of a Model Reporting System and Format.* Washington, D. C.: October 1975.

Glass, Gene V., ed. *Evaluation Studies Review Annual, Volume I.* Beverly Hills, Calif.: Sage Publications, 1976.

Gorham, William, and Glazer, Nathan. *The Urban Predicament.* Washington, D. C.: The Urban Institute, 1976.

Guttentag, Marcia, and Saar, Shalom, eds. *Evaluation Studies Review Annual, Volume 2.* Beverly Hills, Calif.: Sage Publications, 1977.

Hagedorn, Homer J. *Assessment of the Relationship Between Mental Health and Physical Health Planning.* Cambridge: Arthur D. Little, Inc., 1980. (Contract Report HRA 232-79-0090.)

Harris, Deirdre. *Social Services Evaluation: A Review of the Literature.* Working Paper Series 1, no. 1. Austin: Center for Social Work, Graduate School of Social Work, The University of Texas, September 1974.

Hatry, Harry; Blair, Louis; Fisk, Donald; and Kimmel, Wayne. *Program Analysis for State and Local Governments.* Washington, D. C.: The Urban Institute, 1976.

Hatry, Harry P.; Winnie, Richard; and Fisk, Donald M. *Practical Program Evaluation for State and Local Government.* Washington, D. C.: The Urban Institute, 1973.

HEW Evaluation Documentation Center. *Compendium of HEW Evaluation Studies.* Washington, D. C.: Office of the Assistant Secretary for Planning and Evaluation, DHEW, Summer 1980.

Hill, Paul T. "Evaluating Education Programs for Federal Policymakers: Lessons From the NIE Compensatory Education Study." In *Educational Evaluation in the Public Policy Setting*, edited by John Pincus. Santa Monica, Calif.: Rand Corporation, May 1980.

Institute for Advanced Urban Studies. *Final Report: Documentation of Evaluation Efforts in Human Services Agencies in Region VIII.* Denver: University of Colorado, April 15, 1977.

Judson, Horace Freeland. *Eighth Day of Creation: The Makers of the Revolution in Biology.* New York: Simon and Schuster, 1979.

Kimmel, Wayne A. *Needs Assessment: A Critical Perspective.* Washington, D. C.: Office of Program Systems, Office of the Assistant Secretary for Planning and Evaluation, DHEW, December 1977.

Kimmel, Wayne A.; Dougan, William R.; and Hall, John R. *Municipal Management and Budget Methods: An Evaluation of Policy Related Research, Final Report. Volume I: Summary and Synthesis. Volume II: Literature Reviews.* Washington, D. C.: The Urban Institute, 1974.

Knezo, Genevieve J. *Program Evaluation: Emerging Issues of Possible Legislative Concern Relating to the Conduct and Use of Evaluation in the Congress and the Executive Branch.* Washington, D. C.: Library of Congress, Congressional Research Service, November 1974.

Knowles, John H., ed. *Doing Better and Feeling Worse: Health in the United States.* New York: W. W. Norton & Co., 1977.

Larson, Richard C., et al. *Interim Analysis of 200 Evaluations of Criminal Justice.* Cambridge: Operations Research Center, Massachusetts Institute of Technology, May 1979. (Under Grant 78NI-AX0007.)

Levine, R. A., and Williams, A. P. *Making Evaluation Effective: A Guide.* Santa Monica, Calif.: Rand Corporation, May 1971.

Lindblom, Charles E. *Politics and Markets: The World's Political Economic Systems.* New York: Basic Books, Inc., 1977.

Lindblom, Charles E. *The Intelligence of Democracy.* New York: The Free Press, 1965.

Lindblom, Charles E., and Cohen, David K. *Usable Knowledge: Social Science and Social Problem Solving.* New Haven: Yale University Press, 1979.

Lovell, Catherine H., et al. *Federal and State Mandating on Local Governments: An Exploration of Issues and Facts.* Riverside, Calif.: University of California, Graduate School of Administration, June 1979.

Luria, S. E. *Life: The Unfinished Experiment.* New York: Charles Scribner's Sons, 1973.

McLaughlin, Milbrey W. "Evaluation and Reform: The Case of ESEA Title I." Dissertation, Harvard University, 1973.

McLaughlin, Milbrey Wallin. "Evaluation and Alchemy." In *Educational Evaluation in the Public Policy Setting*, edited by John Pincus. Santa Monica, Calif.: Rand Corporation, May 1980.

McLaughlin, Milbrey Wallin. *Evaluation and Reform: The Elementary and Secondary Education Act of 1965, Title I.* Santa Monica, Calif.: Rand Corporation, 1974.

Meld, Murray B. "The Politics of Evaluation of Social Programs." *Social Work* (July 1974).

Morrill, William A., and Francis, Walton J. "Evaluation From the HEW Perspective." In *Evaluation Management: A Selection of Readings.* Washington, D. C.: Office of Personnel Management, Federal Executive Institute, United States of America, January 1979: 25-41.

Morris, W., ed. *The American Heritage Dictionary of the English Language.* Boston: Houghton Mifflin Company, 1976.

National Institute of Mental Health. *A Manual on State Mental Health Planning.* Washington, D. C.: U.S. Government Printing Office, 1977.

Neustadt, Richard E., and Fineberg, Harvey V. *The Swine Flu Affair: Decision-Making on a Slippery Disease.* Washington, D. C.: U.S. Department of Health, Education, and Welfare, 1978.

Nielsen, Victor G. "Program Evaluation and Project Management." Dissertation, University of Utah, 1972.

Pacific Consultants. *Evaluation: A Survey of State Social Service Evaluation Activities.* Berkeley: Pacific Consultants, February 1977.

Patton, Michael Quinn. *Utilization-Focused Evaluation.* Beverly Hills, Calif.: Sage Publications, 1978.

Pearson, Dorothy W. "An Analysis of ESEA Title III Projects." Dissertation, University of Alabama, 1973.

Pincus, John. "The State of Educational Evaluation: Reflections and a Summary." In *Educational Evaluation in the Public Policy Setting*, edited by John Pincus. Santa Monica, Calif.: Rand Corporation, May 1980.

Pincus, John, ed. *Educational Evaluation in the Public Policy Setting.* Santa Monica, Calif.: Rand Corporation, May 1980.

Polanyi, Michael. *Personal Knowledge: Towards a Post-Critical Philosophy.* New York: Harper Torchbooks, 1964.

Pressman, Jeffrey L., and Wildavsky, Aaron B. *Implementation.* Berkeley: University of California Press, 1973.

Region X, U.S. Department of Health, Education, and Welfare. *Ties That Bind.* Seattle: Region X. DHEW, July 1976.

Rein, Martin. *Social Science and Public Policy.* New York and London: Penguin Books, Ltd., 1976.

Rossi, Peter H., and Williams, Walter, eds. *Evaluating Social Programs.* New York and London: Seminar Press, 1972.

Salasin, Susan. "Linking Knowledge to Social Policy-Making: An Interview With Amitai Etzioni." *Evaluation* Special Issue (1978).

Schmidt, Richard E.; Scanlon, John W.; and Bell, James B. *Evaluability Assessment: Making Public Programs Work Better.* Washington, D. C.: The Urban Institute, 1978. (Contract Report No. 1217-50-01 under DHEW Contract No. 100-77-0028.)

Schmidt, Richard E.; Scanlon, John W.; and Bell, James B. *Evaluability Assessment: Making Public Programs Work Better.* Human Services Monograph Series, no. 14. Rockville, Md.: Project SHARE, November 1979.

Scioli, Frank P. "Problems and Prospects for Policy Evaluation." *Public Administration Review* (January/February 1979): 41-45.

Scriven, Michael. "Maximizing the Power of Causal Investigations: The Modus Operandi Method." In *Evaluation Studies Review Annual, Volume I*, edited by Gene V.Glass. Beverly Hills, Calif.: Sage Publications, 1976.

Selye, Hans. *The Stress of Life.* New York: McGraw Hill Book Co., 1976.

Sharpe, L. J. "The Social Scientist and Policy-Making: Some Cautionary Thoughts and Transatlantic Reflections." *Policy and Politics* (1976).

Sobel, David S., ed. *Ways of Health.* New York: Harcourt Brace Jovanovich, 1979.

Stockdill, James W. "The Politics of Program Evaluation." Paper presented at the Florida Mental Health Evaluation Consortium, May 1974.

Stokey, Edith, and Zeckhauser, Richard. *A Primer for Policy Analysis.* New York: W. W. Norton & Co., 1978.

Suchman, Edward A. *Evaluative Research: Principles and Practice in Public Service and Social Action Programs.* New York: Russell Sage Foundation, 1967.

Thorson, Thomas Landon. *Biopolitics.* New York: Holt, Rinehart & Winston, Inc., 1970.

Toulmin, Stephen. *Human Understanding.* Princeton: Princeton University Press, 1972.

U.S. General Accounting Office. *Improvements Needed in the Administration of Contracts for Evaluations and Studies of Antipoverty Programs.* Washington, D. C.: GAO, December 1971.

Wattenberg, B. J. *In Search of the Real America.* New York: Berkley Publishing Corp., 1978.

Weiss, Carol H. "Between the Cup and the Lip." *Evaluation* (1973).

Weiss, Carol H. *Evaluation Research: Methods of Assessing Program Effectiveness.* Englewood Cliffs, N.J.: Prentice Hall, Inc., 1972.

Weiss, Carol H. "Research for Policy's Sake: The Enlightenment Function of Social Research." *Policy Analysis* 3, no. 4 (1977): 531-545.

Weiss, Carol H., ed. "Symposium on 'The Research Utilization Quandary.' " *Policy Studies Journal* (Spring 1976).

White, Bayla F. *The Atlanta Project: How One Large School System Responded to Performance Evaluation.* Washington, D. C.: The Urban Institute, March 1977.

Wholey, Joseph S. "Contributions of Social Intervention Research to Government Practices." *Annals of the New York Academy of Sciences* 218 (1973).

Wholey, Joseph S. *Evaluation: Promise and Performance.* Washington, D. C.: The Urban Institute, 1979.

Wholey, Joseph S. *Zero-Base Budgeting and Program Evaluation.* Lexington and Toronto: Lexington Books, 1978.

Wholey, Joseph S., et al. *Federal Evaluation Policy: Analyzing the Effects of Public Programs.* Washington, D. C.: The Urban Institute, 1976.

Wildavsky, Aaron V. *Speaking Truth to Power: The Art and Craft of Policy Analysis.* Boston: Little, Brown & Co., 1979.

Williams, Walter, and Rossi, Peter H. *Evaluating Social Programs.* New York: Seminar Press, 1972.

Windle, C. "A Crisis for Program Evaluation: An Embarrassment of Opportunity." *Rhode Island Medical Journal* (November 1976).

Windle, Charles; Bass, Rosalyn D.; and Taube, Carl A. "PR Aside: Initial Results from NIMH's Service Program Evaluation Studies." *American Journal of Community Psychology* 2, no. 3 (1974).

Windle, Charles, and Neigher, William. "Ethical Problems in Program Evaluation: Advice for Trapped Evaluators." *Evaluation and Program Planning* 1 (1978): 97-108.

Zangwill, Bruce. *A Compendium of Laws and Regulations Requiring Needs Assessment.* Washington, D. C.: Office of Program Systems, OASPE, OS, DHEW, May 1977.

*U.S. GOVERNMENT PRINTING OFFICE: 1981-0-341-155/108